Universidad
Carlos III de Madrid

TESIS DOCTORAL

Essays in High Dimensional Factor Models

Autor:
Liang Chen

Director/es:
Juan José Dolado y Jesús Gonzalo

DEPARTAMENTO DE ECONOMÍA

Getafe, julio de 2013

TESIS DOCTORAL

Essays in High Dimensional Factor Models

Autor: Liang Chen

Director/es: Juan José Dolado y Jesús Gonzalo

Firma del Tribunal Calificador:

Firma

Presidente:

Vocal:

Secretario:

Calificación:

Getafe, de de

Abstract

My PhD thesis consists of three chapters on high dimensional factor models and their applications. In Chapter 1, I study how to test for structural breaks in large factor models. Time invariance of factor loadings is a standard assumption in the analysis of large factor models. Yet, this assumption may be restrictive unless parameter shifts are mild. In this chapter we develop a new testing procedure to detect *big* breaks in these loadings at either known or unknown dates. The test fares well in terms of power relative to other recently proposed tests on this issue, and can be easily implemented to avoid forecasting failures in standard factor-augmented models where the number of factors is a priori imposed on the basis of theoretical considerations.

Despite their growing popularity, factor models have been often criticized for lack of identification of the factors. In Chapter 2, I try to identify the orthogonal factors estimated using principal component by associating them to a relevant subset of observed variables. I first propose a selection procedure to choose such a subset, and then test the hypothesis that true factors are exact linear combinations of the selected variables. The good performance of my method in finite samples and its advantages relative to the other available procedures are confirmed through simulations. Empirical applications include the identification of the underlying risk factors in large dataset of stock and portfolio returns, as well as interpreting the factors in a large panel of macroeconomic time series. In both cases, it is shown that the underlying factors can be closely approximated by a few observed variables.

In Chapter 3 I investigate the source of the aggregate volatility in industrial productions (IP) using factor models. I consider 3 structural dynamic macro models with multiple producing sectors. General conditions are given to show how the sectoral IP growth rates can be represented as a dynamic factor model (DFM) through input-output linkages. Using available data, we first investigate whether the input-output linkages in these models are strong enough to generate a DFM representation for the sectoral IP growth rates. We also find that after the great moderation in 1984, the sectoral IP growth rates can be characterized by an approximate factor model with only 1 common factor, which is found to be connected primarily to a aggregate technology shock that affects most of the sectors, and possibly to 1 or 2 sectoral shocks that only affect the key sectors that provide inputs for many other sectors.

Resumen

Mi tesis consta de tres capítulos sobre modelos de factores de alta dimensión y sus aplicaciones. En el capítulo 1, se investiga cómo contrastar los cambios estructurales en los modelos de factores de alta dimensión. Invariancia del Tiempo de carga factorial es un supuesto estndar en el análisis de modelos de factores de alta dimensión. Sin embargo, esta hipótesis puede ser restrictiva a menos que cambios de parámetros no sean grandes. En este capítulo se desarrolla un nuevo procedimiento para detectar grandes roturas en estas cargas en fechas previamente conocidas o no. El contraste propuesto, una vez comparado con otras propuestas similares, presenta una buena performance en termos de poder. Además, puede ser fácilmente implementado para evitar fallos de previsión en modelos aumentado de factores estándares, donde el número de factores es impuesto segundo consideraciones teóricas.

A pesar de su creciente popularidad, los modelos de factores han sido criticados por la falta de identificación de los factores. En el capítulo 2, intento identificar los factores ortogonales estimados utilizando componentes principales asociándolos a un subconjunto relevante de variables observadas. Primeramente se propone un procedimiento de selección para elegir un subconjunto y, a continuación, contrastar la hipótesis de que los factores reales son combinaciones lineales exactas de las variables seleccionadas. El buen rendimiento de mi método en muestras finitas y sus ventajas en relación con los otros procedimientos disponibles se confirman a través de simulaciones. Aplicaciones empíricas incluyen la identificación de los factores de riesgo subyacentes en la gran base de datos de la cartera de valores y rendimientos, así como la interpretación de los factores en un gran panel de series temporales macroeconómicas. En ambos casos, se muestra que los factores subyacentes pueden ser estrechamente aproximados por unas pocas variables observadas.

En el capítulo 3 investigo el origen de la volatilidad agregada en las producciones industriales (IP) mediante modelos de factores. Considero 3 modelos macroeconómicos dinámicos estructurales con múltiples sectores productivos. Se dan las condiciones generales de mostrar cómo las tasas de crecimiento sectoriales IP se pueden representar como un modelo de factores dinámicos (DFM) a través de enlaces de entrada y salida. Utilizando los datos disponibles, investigamos si los vínculos de entrada y salida en estos modelos son lo suficientemente fuertes como para generar una representación DFM de las tasas de crecimiento sectoriales IP. También encontramos que después de la gran moderación en 1984, las tasas de crecimiento sectoriales IP pueden ser caracterizadas por un modelo aproximado factores con sólo 1 factor común, que se encuentra a conectar todo a un shock tecnológico global que afecta a la mayoría de los sectores, y posiblemente a 1 o 2 choques sectoriales que sólo afectan a los sectores clave que proveen insumos para muchos otros sectores.

Acknowledgements

First of all, my deepest thanks go to my supervisors *Juan J. Dolado* and *Jesús Gonzalo* for suggesting the topic of the thesis, as well as for their continuous support, advice and encouragement.

Likewise, I would like to express my gratitude to my thesis committee members: Carlos Velasco, Abderrahim Taamouti and Alfonso Valdesogo for there insightful comments and constructive critique.

Finally, I also would like to thank the following faculty and graduate students in the PhD program in economics at UC3M for their helpful discussions and suggestions in several internal seminars: Prof. Miguel Delgado, Prof. Noelia Cámara, Yunus Emre, Mian Huang, Omar Rachedi, Fabian Rinnen, Francesco Risi, Pedro Sant'Anna and Xiaojun Song.

All errors are mine.

Contents

Abstract i

Resumen ii

Acknowledgements iii

List of Figures vi

List of Tables vii

1 Detecting Big Structural Break in Large Factor Models 1
- 1.1 Introduction . . . 1
- 1.2 Notation and Preliminaries . . . 4
- 1.3 The Effects of Structural Breaks . . . 6
  - 1.3.1 The estimation of factors . . . 7
  - 1.3.2 The estimated number of factors . . . 10
- 1.4 Testing for Structural Breaks . . . 11
  - 1.4.1 Hypotheses of interest and test statistics . . . 11
  - 1.4.2 Limiting distributions under the null hypothesis . . . 13
  - 1.4.3 Performance of the tests under the alternative hypothesis . . . 15
  - 1.4.4 Disentangling breaks in loadings from breaks in factors . . . 17
- 1.5 Simulations . . . 18
  - 1.5.1 The effect of big breaks on forecasting . . . 18
  - 1.5.2 Size properties . . . 19
  - 1.5.3 Power properties . . . 24
  - 1.5.4 Comparison with the BE test . . . 26
  - 1.5.5 Comparison with the HI test . . . 28
- 1.6 An Empirical Application . . . 29
- 1.7 Conclusions . . . 30

2 Identifying Observed Factors in High Dimensional Factor Models 33
- 2.1 Introduction . . . 33
- 2.2 Models, Notations and Assumptions . . . 36
- 2.3 Identifying Observed Factors Using Regressions . . . 38
  - 2.3.1 Directly Observed Factos . . . 38
  - 2.3.2 Indirectly Observed Factors . . . 41

2.3.2.1 Definitions and comparison with Bai and Ng (2006) 41
2.3.2.2 Identifying the IOFs 42
2.3.2.3 Practical implementation 44
2.3.3 Identification Based on the Lasso 45
2.3.3.1 Identification of observed factors using the adaptive Lasso . . 45
2.3.3.2 Computations and others 48
2.3.4 Weakly Influential Factors 49
2.4 Hypothesis Testing 51
2.5 Simulations 55
2.5.1 Directly Observed Factors 55
2.5.2 Indirectly Observed Factors 57
2.5.3 The Lasso and the Adaptive Lasso 58
2.6 Applications 60
2.6.1 Factors in Portfolio Returns 60
2.6.2 Macroeconomic Factors 63
2.6.3 Factors in Stock Market 65
2.7 Conclusion 68

3 The Source of the Aggregate Volatility in Industrial Productions 73
3.1 Introduction 73
3.2 Sector-Specific Shocks, Input-Output Linkages, and Dynamic Factor Models . 75
3.2.1 Notations and structural models 75
3.2.2 Input-output linkages and dynamic factor models 78
3.2.3 Discussion using a simple example 80
3.3 Empirical Analysis of Sectoral IP Growth Rates Using Factor Models 82
3.3.1 Can sector-specific shocks generate aggregate volatility? 82
3.3.2 Number of factors and structural breaks 83
3.3.3 Identifying the common factor 86
3.4 Conclusion 91

A Appendix to Chapter 1 92
A.1 Proof of Propositions 1.1 and 1.2 92
A.2 Proof of Theorem 1.3 93
A.3 Consistent Estimator of S 97

B Appendix to Chapter 2 100
B.1 Proof of Theorem 2.2 100
B.2 Proof of Theorem 2.3 102
B.3 Proof of Theorem 2.4 105
B.4 Proof of Theorem 2.6 109
B.5 Tables and Figures 112

Bibliography 114

List of Figures

1.1 The MSEs of different forecasting methods in the presence of big breaks (see Section 1.5.1). 22
1.2 US data set of Stock and Watson (2009), from 1959:Q1 to 2006:Q4. The trimming $\Pi = [0.3, 0.7]$ is used for the Wald tests with $\bar{r} = 2$ to 6 (from top to bottom), and the horizontal lines are the corresponding 5% asymptotic critical values for the Sup-Wald Test. 31

2.1 R^2 in the regressions of the variables listed on the x axis onto the estimated factors. 66
2.2 The estimated factors and the fitted values in the regressions 1 and 2. 67

3.1 Long and Plosser 83
3.2 Horvath 83
3.3 FSW 83
3.4 The Evidence of One Static Factor 86
3.5 Distribution of $\sum_{j=1}^{N} \gamma_{ij}$ 88
3.6 Distribution of $\sum_{j=1}^{N} \mathbf{1}(\gamma_{ij} > 0)$ 88

List of Tables

1.1 Empirical Sizes of the Nominal 5% Size Tests for Different Choices of $\overline{r}$ when $r = 3$. . . . 20
1.2 Empirical Size of the Nominal 5% Size Tests for Different Choices of $\bar{r}$ when $r = 3$, and when the idiosyncratic errors are cross sectionally and serially correlated. . . . 23
1.3 Empirical Power of Nominal 5% Size Tests for Different Choices of $\bar{r}$ when $r = 2$ and $k_1 = 2$. . . . 24
1.4 Empirical Power of Nominal 5% Size Tests for Different Choices of $\bar{r}$ when $r = 2$, $k_1 = 1$, and $\hat{F}_{1t}$ is used as the regressand. . . . 25
1.5 Empirical Power of Nominal 5% Size Tests for Different Choices of $\bar{r}$ when $r = 2$, $k_1 = 1$, and $\hat{F}_{2t}$ is used as the regressand. . . . 26
1.6 Size and Power Comparisons of BE (2011) and Our Wald Tests at Nominal 5% Size for $r = 2$. . . . 27
1.7 Power Comparison of HI (2012) and Our Wald Tests for $N, T \leq 100$ when $r = 1$. 29

2.1 Probabilities of Correctly Identifying DOFs. . . . 56
2.2 Test with DOFs 57
2.3 Probabilities of Correctly Identifying IOFs 58
2.4 Test with IOFs 59
2.5 Identifying DOF using Lasso and adaptive Lasso. . . . 69
2.6 Identifying IOF using Lasso and adaptive Lasso. . . . 69
2.7 The estimated number of factors using the information criteria of Bai and Ng (2002) (PC_i and IC_i) and the method of Onatski (2010), with $r_{max} = 10$. . . . 69
2.8 Identification of observed factors for the returns of portfolios 70
2.9 Regressions of estimated factors on observed factors. . . . 70
2.10 The estimated number of factors using the information criteria of Bai and Ng (2002) and the method of Onatski (2010), with $r_{max} = 10$, for band pass filtered macro data sets from 1964 to 2008. . . . 71
2.11 Identification of observed factors for the returns of portfolios 71
2.12 Details of the selected observed factors 71
2.13 Selected observed factors for the stock returns using $\tilde{f}_{1t}$ 71
2.14 Selected observed factors for the stock returns using $\tilde{f}_{2t}$ 72

3.1 The Estimated Number of Static Factors 85
3.2 Testing the Number of Static Factors 85
3.3 The Estimated Number of Static Factors for $\hat{u}_t$ 87
3.4 Top 10 Sectors Ranked by $\sum_{j=1}^{N} \gamma_{ij}$ 88
3.5 Top 10 Sectors Ranked by $\sum_{j=1}^{N} \mathbf{1}(\gamma_{ij} > 0)$ 88
3.6 Identified Observed Factors Using Structural Models 90

B.1 Candidates for Observed Factors 113

Chapter 1

Detecting Big Structural Break in Large Factor Models

1.1 Introduction

Despite being well acknowledged that some parameters in economic relationships can become unstable due to important structural breaks (e.g., those related to technological change, globalization or strong policy reforms), a standard practice in the estimation of large factor models (FM, hereafter) has been to assume time-invariant factor loadings. Possibly, one of the main reasons for this benign neglect of breaks stems from the important results obtained by Stock and Watson (2002a, 2009) regarding the consistency of the estimated factors by principal components analysis (PCA hereafter) when the loadings are subject to small (i.e., local-to-zero) instabilities. These authors conclude that the failure of factor-based forecasts is mainly due to the instability of the forecast function, rather than of the different components of the FM. As a result, their advice is to use full-sample factor estimates and subsample forecasting equations to improve forecasts.

However, the main emphasis placed on local-to-zero breaks has been subsequently questioned. For example, by means of a Monte Carlo study, Banerjee, Marcellino and Masten (2008) conclude that, in contrast to Stock and Watson's diagnosis, the instability of factor loadings when big (i.e., not local-to-zero) breaks occur is the most likely reason behind the worsening factor-based forecasts, particularly in small samples. Likewise, when discussing Stock and Watson's research on this topic, Giannone (2007) argues that *"....to understand structural changes we should devote more effort in modelling the variables characterized by more severe instabilities..."*. In this paper, we pursue this goal by providing a precise characterization of the different conditions under which big and small breaks in the factor loadings may occur, as well as develop a simple test to distinguish between them. We conclude that, in contrast to small breaks, big breaks should not be ignored in our setup since they may lead to misleading

results in standard econometric practices using FM and in the potential interpretation of the factors themselves.

A forerunner of our paper is Breitung and Eickmeier (2011, BE henceforth) who were the first to propose a proper testing procedure to detect big breaks in the factor loadings. Their test relies on the idea that, under the null of no structural break (plus some additional assumptions), the estimation error of the factors can be ignored and thus the estimated factors can be treated as the true ones. Consequently, a Chow-type test can be implemented by regressing each variable in the data set on both the estimated factors using the whole sample and their truncated versions from the date of the break onwards. Focusing on the joint statistical significance of the estimated coefficients on the truncated factors, their test compares the empirical rejection frequency among the individual regressions to a nominal size of 5% under the null. In our view, this approach suffers from two limitations: (i) the overall limiting distribution of their test remains unknown when testing for the equality of all the elements of the loading matrix in both subsamples;[1] and (ii) it lacks non-trivial power when the number of factors is chosen according to some consistent estimator of r. This last problem can be serious. For example, as explained further below, a FM with r original factors where the loadings of one of them exhibit a big structural break at the same date admits a standard factor representation with $r + 1$ factors without a break. Hence, if the number of factors were to be chosen as $r + 1$, instead of r, their testing approach may not detect any break at all when in fact there is one.

Our contribution here is to propose a simple testing procedure to detect big breaks in FMs stemming from a single source which does not suffer from the previous shortcomings. We focus on breaks in the loadings though we also provide a procedure to detect whether the breaks originate from the loadings or from factors themselves. In particular, we first derive some asymptotic results finding that, in contrast to small breaks, the number of factors is overestimated under big breaks, a result which was also used by BE (2011). We argue that neglecting these breaks can have serious consequences on the forecasting performance of some popular regression-based models using factors, where their number is a priori imposed. Likewise, under big breaks, it may be difficult to provide a structural interpretation of the estimated factors when they are chosen according to some consistent information criteria (see Bai and Ng, 2006b, and Chen, 2012). Our proposal relies upon a very simple regression-based testing procedure. As sketched earlier, the insight is that a FM with big breaks in the loadings can be re-parameterized as another FM with constant loadings but a larger set of factors, where the number and the space spanned by the latter can be consistently estimated by PCA under fairly standard assumptions. Hence, rather than directly testing for whether all the elements of the loadings matrix are stable, which will suffer from an infinite-dimensionality problem as the number of variables in the panel data set grows, one

[1] With the notation used below in (1.1)- (1.2), the limiting distribution of the rejection frequencies for the joint hypothesis $A = B$ is not known, although the individual tests for the hypothesis $\alpha_i = \beta_i$ have known limiting distributions.

can test if the relationships among the larger finite-dimensional set of estimated factors are stable.

Specifically, our procedure consists of two steps. First, the number of factors for the whole sample period is chosen as $\bar{r}$ according to Bai and Ng's (2002; BN henceforth) information criteria, and then $\bar{r}$ factors are estimated by PCA. Next, one of the estimated factors (e.g., the first one) is regressed on the remaining $\bar{r}-1$ factors, to next apply the standard Chow Test or the Sup-type Test of Andrews (1993) to this regression, depending on whether the date of the break is treated as known or unknown. If the null of no breaks is rejected in the second-step regression, we conclude that there are big breaks and, otherwise, that either no breaks exist at all or that only small breaks occur. Further, on the basis of the rank properties of the covariance matrix of the estimated factors in different subsamples, we also provide some guidance on how to distinguish between breaks stemming either from the loadings or from the data generating process (DGP hereafter) of the factors. This difference is important since the latter may lead to reject the null of constant loadings when it is true, implying a misleading interpretation of the source of the break.

After completing the first draft of this paper, we became aware of a closely related paper by Han and Inoue (2012, HI hereafter) who, in an independent research, adopt a similar approach to ours in testing for big breaks in FM. The two approaches, however, differ in some relevant respects. In effect, rather than using a simple regression-based approach to avoid the infinite-dimensionality problem, as we do here, HI (2012) test directly for differences before and after the break in all the elements of the covariance matrix of the estimated factors. We will argue below that, despite the fact that the HI tests use more information than ours, our tests generally exhibit similar power. Indeed, our regression-based test based on the Wald principle, which behaves much better in general than the Lagrange multiplier (LM hereafter) tests for detecting structural breaks, is even more powerful than the corresponding HI's test for small sample sizes, such as $N = T = 50$. One additional advantage of our simple linear-regression setup is that it amenable to use many other existing methods for testing breaks, including multiple ones (see, e.g., Perron, 2006, for an extensive review of these tests).

The rest of the paper is organized as follows. In Section 1.2, we present the basic notation, assumptions and the definitions of *small* and *big* breaks. In Section 1.3, we analyze the consequences of big breaks on the choice of the number of factors and their estimation, as well as their effects on standard econometric practices with factor-augmented regressions. In Section 1.4, we first derive the asymptotic distributions of our tests and next discuss, when a big break is detected, how one can identify whether it stems from the loadings or from the process driving the factors. Section 1.5 deals with the finite sample performance of our test relative to the competing tests using Monte-Carlo simulations. Section 1.6 provides an empirical application. Finally, Section 7 concludes. An Appendix contains detailed proofs of the main results.

1.2 Notation and Preliminaries

We consider FM that can be rewritten in the static canonical form:

$$X_t = AF_t + e_t$$

where X_t is the $N \times 1$ vector of observed variables, $A = (\alpha_1, \ldots, \alpha_N)'$ is the $N \times r$ matrix of factor loadings, r is the number of common factors which is finite, $F_t = (F_{1t}, \ldots, F_{rt})'$ is the $r \times 1$ vector of common factors, and e_t is the $N \times 1$ vector of idiosyncratic errors. In the case of dynamic FMs, all the common factors f_t and their lags are stacked into F_t. Thus, a dynamic FM with r dynamic factors and p lags of these factors can be written as a static FM with $r \times (p+1)$ static factors. Further, given the assumptions we make about e_t, the case analyzed by BE (2011) where the e_{it} disturbances are generated by individual specific autoregressive (AR hereafter) processes is also considered.[2]

We assume that there is a single structural break in the factor loadings of all factors at the same date τ:

$$X_t = AF_t + e_t \quad t = 1, 2, \ldots, \tau, \tag{1.1}$$

$$X_t = BF_t + e_t \quad t = \tau + 1, \ldots, T \tag{1.2}$$

where $B = (\beta_1, \ldots, \beta_N)'$ is the new factor loadings after the break. By defining the matrix $C = B - A$, which captures the size of the breaks, the FM in (1.1) and (1.2) can be rewritten as:

$$X_t = AF_t + CG_t + e_t \tag{1.3}$$

where $G_t = 0$ for $t = 1, \ldots, \tau$, and $G_t = F_t$ for $t = \tau + 1, \ldots, T$.

As argued by Stock and Watson (2002, 2009), the effects of some mild (local to zero) instability in the factor loadings can be averaged out, so that estimation and inference based on PCA remain valid. We generalize their analysis by allowing for two types of break sizes: *small* and *big*. In contrast to the former, we will show that the latter cannot be neglected. To distinguish between them, it is convenient to partition the C matrix as follows:

$$C = [\Lambda \quad H]$$

where Λ and H are $N \times k_1$ and $N \times k_2$ matrices that correspond to the *big* and the *small* breaks, and $k_1 + k_2 = r$. Accordingly, we can also partition the G_t matrix into G_t^1 and G_t^2, such that (1.3) becomes:

$$X_t = AF_t + \Lambda G_t^1 + HG_t^2 + e_t \tag{1.4}$$

where $\Lambda = (\lambda_1, \ldots, \lambda_N)'$ and $H = (\eta_1, \ldots, \eta_N)'$.

[2] Notice, however, that our setup excludes the generalized dynamic FM considered by Forni and Lippi (2001), where the polynomial distributed lag possibly tends to infinity.

Throughout the paper, $tr(\Sigma)$ and $||\Sigma|| = \sqrt{tr(\Sigma'\Sigma)}$ will denote the trace and the norm of a matrix Σ, respectively. For a finite dimensional vector v, we write $v = O_p(1)$ ($v = o_p(1)$) when $\|v\| = O_p(1)$ ($\|v\| = o_p(1)$). $[T\pi]$ denotes the integer part of $T \times \pi$ for $\pi \in [0, 1]$. Once the basic notation has been established, the next step is to provide precise definitions of the two types of breaks.

Assumption 1. ***Breaks:*** *(a)* $||\lambda_i|| \leq \bar{\lambda} < \infty$ *for all* i. $N^{-1}\Lambda'\Lambda \to \Sigma_\Lambda$ *as* $N \to \infty$ *for some positive definite matrix* Σ_Λ. *(b)* $\eta_i = (NT)^{-1/2}\kappa_i$ *and* $\|\kappa_i\| \leq \bar{\kappa} < \infty$ *for all* i.

The matrices Λ and H are assumed to contain non-random elements. Assumption 1.a yields the definition of a big break. It also includes the case where $\lambda_i = 0$ (no break) for a fixed proportion of variables as $N \to \infty$. As will be shown, this type of breaks will lead to inconsistency of the estimated factors and the overestimation of r. Assumption 1.b, in turn, provides the definition of small breaks which are characterized as being of order $1/\sqrt{NT}$, so the true factor space and r can be both consistently estimated under such breaks. If we only focus on the consistency of the estimated factors, the definition of small breaks can be relaxed to $\|\eta_i\| = O(N^{\delta_1}T^{\delta_2})$ with $\delta_1, \delta_2 \leq 0$ and $\delta_1 + \delta_2 \neq 0$. As shown by Theorem 1 of Bates et al. (2013), the estimated factors using PCA are still consistent for breaks with such sizes, although at a slower convergence rate than $\min(\sqrt{N}, \sqrt{T})$. However, as argued in the Introduction, we are concerned with the number of factors as well, and such breaks will possibly lead to overestimated number of factors. Therefore, our assumption for small breaks is more stringent.

Remark 1. Bates et al. (2013) considers instabilities in the factor loadings such that: $A_t = A + h_{NT} * u_t$, where h_{NT} is a scalar which depends on N and T, and u_t is a vector of possibly random disturbances. They propose conditions about h_{NT} and u_t under which the PCA estimator of the factors are still consistent. In the case of structural breaks, they assume $h_{NT} = 1$ whereas u_t is vector of $O(1)$ elements (say $u_t = \mathbf{\Delta 1}(t > \tau)$ where $\Delta = (\Delta_1, \ldots, \Delta_N)'$) that do not depend on N or T after the break date τ. Unlike our way of defining the break sizes, they characterize the *size* of breaks as the number of nonzero elements in Δ, i.e., the number of variables having breaks in their factor loadings. For the consistency of the estimated factors, their conditions allow at most N^δ ($\delta < 1$) variables to have breaks ($\sum_{i=1}^N ||\Delta_i|| = O(N^\delta)$). By contrast, for the consistency of the estimated *number* of factors using BN's (2002) method, only a fixed number of variables are allowed to have breaks ($\sum_{i=1}^N ||\Delta_i|| = O(1)$), when N/T converges to some nonzero constant.

To compare these conditions with our definitions, notice that when N^δ variables are allowed to have breaks, we have $\left\|N^{-1}\sum_{i=1}^N \Delta_i\Delta_i'\right\| \leq N^{-1}\sum_{i=1}^N \|\Delta_i\|^2 = O(N^{\delta-1}) = o(1)$ when $\delta < 1$. On the other hand, our Assumption 1.a for big breaks implies that $\left\|\frac{1}{N}\sum_{i=1}^N \lambda_i\lambda_i'\right\|$ converges to a positive constant. In this sense, the big breaks defined in our paper have larger sizes than those considered in Bates et al. (2013), under which the estimated factors are proved to be consistent. As for the small breaks, our Assumption 1.b implies that

$\sum_{i=1}^{N} ||\eta_i|| = O(1)$ when N and T have the same order, similar to the conditions of Bates et al. (2003) which also allows the number of factors to be consistently estimated. □

To investigate the influence of the breaks on the number and estimation of factors, some further assumptions need to be imposed. To achieve consistent notation with the previous literature in the discussion of these assumptions, we follow the presentation of BN (2002) and Bai (2003) with a few slight modifications.

Assumption 2. ***Factors:*** *$E(F_t) = 0$, $E||F_t||^4 < \infty$, $(NT)^{-4}E\|F_t\|^8 < \infty$, $T^{-1}\sum_{t=1}^{T} F_t F_t' \xrightarrow{p} \Sigma_F$ and $T^{-1}\sum_{t=1}^{\tau} F_t F_t' \xrightarrow{p} \pi^*\Sigma_F$ as $T \to \infty$ for some positive definite matrix Σ_F where $\pi^* = \lim_{T\to\infty} \frac{\tau}{T}$.*

Assumption 3. ***Factor Loadings:*** *$||\alpha_i|| \leq \bar{\alpha} < \infty$, and $N^{-1}A'A \to \Sigma_A$, $N^{-1}\Gamma'\Gamma \to \Sigma_\Gamma$ as $N \to \infty$ for some positive definite matrix Σ_A and Σ_Γ, where $\Gamma = [A \quad \Lambda]$.*

Assumption 4. ***Idiosyncratic Errors:*** *The error terms e_t, the factors F_t and the loadings A satisfy the Assumption A, B, C, D, E, F1 and F2 of Bai (2003).*

Assumption 5. ***Independence of Factors and Idiosyncratic Errors:*** *$[F_t]_{t=1}^{T}$ and $[e_t]_{t=1}^{T}$ are two mutually independent groups, and $1/\sqrt{T}\sum_{t=1}^{T} F_t e_{it} = O_p(1)$ for each $i = 1, \ldots, N$.*

Assumptions 3 and 4 are standard in the literature on FM allowing for weak cross-sectional and temporal correlations between the errors. Notice that, in our specific setup, Assumption 3 excludes the case where a new (old) factor appears (disappears) after the break since this event would imply that Σ_Γ becomes singular. However, this is not restrictive since we could always envisage any potential factor as having non-zero, albeit small, loadings in either of the relevant subsamples. Assumption 2, in turn, is a new one. Since factors and factor loadings cannot be separately identified, we have to assume that DGPs with breaks in the loadings, which can be reabsorbed by transformations of the factors, should not be included in the alternative. In Section 1.4.4, we will discuss how to differentiate between breaks in the factor loadings and breaks in the dynamics of the factors. Different factors are allowed to be correlated at all leads and lags. Assumption 5 on the independence among the different groups is stronger than the usual assumptions made by BN (2002). Notice, however, that we could have also assumed some dependence between these groups and then impose some restrictions on this dependence when necessary. Yet, this would complicate the proofs without essentially altering the insight underlying our approach. Thus, for the sake of simplicity, we assume them to be independent in the sequel.

1.3 The Effects of Structural Breaks

In this section, we study the effects of structural breaks on the estimation of both the number of factors based on the information criteria (IC, henceforth) proposed by BN (2002) and the

factors themselves through PCA. Our main finding is that, in contrast to Stock and Watson's (2002, 2009) consistency result for the true factor space under small breaks, the factor space estimated by PCA is inconsistent, and that the number of factors tends to be overestimated under big breaks.

1.3.1 The estimation of factors

Let us rewrite model (1.4) with k_1 big breaks and k_2 small breaks in the more compact form:

$$X_t = AF_t + \Lambda G_t^1 + \epsilon_t \tag{1.5}$$

where $\epsilon_t = HG_t^2 + e_t$. The idea is to show that the new error terms ϵ_t still satisfy the necessary conditions for (1.5) being a standard FM with new factors $F_t^* = [F_t' \quad G_t^{1'}]'$ and new factor loadings $[A \quad \Lambda]$.

Let $\bar{r}$ be the selected number of factors, either by some prior knowledge or using some consistent estimator such as the IC of BN (2002), where notice that $\bar{r}$ is not necessarily equal to r. Let $\hat{F}$ be $\sqrt{T}$ times the $\bar{r}$ eigenvectors corresponding to the $\bar{r}$ largest eigenvalues of the matrix XX', where the $T \times N$ matrix $X = [\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_T]'$, $\bar{X}_t = [X_{t1}, X_{t2}, \ldots, X_{tN}]'$, $\hat{F} = [\hat{F}_1, \hat{F}_2, \ldots, \hat{F}_T]'$. Then we have:

Proposition 1.1. *For any fixed $\bar{r}$ ($1 \leq \bar{r} \leq r + k_1$), under Assumptions 1 to 5, there exists a full rank $\bar{r} \times (r + k_1)$ matrix D and $\delta_{N,T} = \min\{\sqrt{N}, \sqrt{T}\}$ such that:*

$$\hat{F}_t = DF_t^* + O_p(\delta_{N,T}^{-1}) \text{ for } t = 1, 2, \ldots, T. \tag{1.6}$$

This result implies that $\hat{F}_t$ estimate consistently the space of the new factors, F_t^*, but not the space of the original factors, F_t.

Let us next consider two cases. First, when $k_1 = 0$ (no big breaks), we have that $G_t^1 = 0$, and $F_t^* = F_t$, so that (1.6) becomes

$$\hat{F}_t = DF_t + O_p(\delta_{N,T}^{-1}) \tag{1.7}$$

for a $\bar{r} \times r$ matrix D of full rank. This just trivially replicates the well-known consistency result of BN (2002).[3]

Secondly, in the more interesting case where $k_1 > 0$ (big breaks exist), we can rewrite (1.6) as

$$\hat{F}_t = [D_1 \quad D_2] \begin{pmatrix} F_t \\ G_t^1 \end{pmatrix} + o_p(1) = D_1 F_t + D_2 G_t^1 + o_p(1) \tag{1.8}$$

[3]Notice that, for the estimator $\hat{F}$ defined here, $\bar{r}$ has to be smaller or equal to r for (1.7) to hold.

where the $\bar{r} \times (r + k_1)$ matrix D is partitioned into the $\bar{r} \times r$ matrix D_1 and the $\bar{r} \times k_1$ matrix D_2. Note that, by the definition of G_t, $G_t^1 = 0$ for $t = 1, 2, \ldots, \tau$, and $G_t^1 = F_t^1$ for $t = \tau + 1, \ldots, T$, where F_t^1 is the $k_1 \times 1$ sub-vector of F_t that is subject to big breaks in their loadings. Therefore (1.8) can be expressed as:

$$\hat{F}_t = D_1 F_t + o_p(1) \text{ for } t = 1, 2, \ldots, \tau, \tag{1.9}$$

$$\hat{F}_t = D_2^{\dagger} F_t + o_p(1) \text{ for } t = \tau + 1, \ldots, T \tag{1.10}$$

where $D_2^{\dagger} = D_1 + [D_2 \quad 0]$, 0 is a $\bar{r} \times (r - k_1)$ zero matrix, and in general $D_2 \neq 0$. Hence, since $D_1 \neq D_2^{\dagger}$, this implies that, in contrast to small breaks where D_2 tends to 0 due to the local-to-zero properties of the elements of H (see Assumption 1.b), under big breaks the estimated factors $\hat{F}$ will not be consistent for the space of the true factors F. Accordingly, as will be explained below, imposing a priori the number of estimated factors to be used as predictors or explanatory variables in standard factor-augmented models may lead to misleading results.

To illustrate the consequences of having big breaks in the factor loadings, consider the following simple Factor Augmented Regression (FAR) model (see BN, 2006a):

$$y_t = a' F_t + b' W_t + u_t, \ t = 1, 2, \ldots, T \tag{1.11}$$

where W_t is a small set of observable variables and the $r \times 1$ vector F_t contains the r common factors driving a large panel dataset X_{it} $(i = 1, 2, \ldots, N;\ t = 1, 2, \ldots, T)$ which excludes both y_t and W_t. The parameters of interest are the elements of vector b while F_t is included in (1.11) to control for potential endogeneity arising from omitted variables. Since we cannot identify F_t and a, only the product $a' F_t$ is relevant. Suppose that there is a big break at date τ. From (1.9) and (1.10), we can rewrite (1.11) as:

$$\begin{aligned} y_t &= (a' D_1^{-})(D_1 F_t) + b' W_t + u_t \ \text{ for } t = 1, 2, \ldots, \tau, \\ y_t &= (a' D_2^{\dagger -})(D_2^{\dagger} F_t) + b' W_t + u_t \text{ for } t = \tau + 1, \ldots, T \end{aligned}$$

where $D_1^{-} D_1 = D_2^{\dagger -} D_2 = I_r$, or equivalently

$$y_t = a_1' \hat{F}_t + b' W_t + \tilde{u}_t \ \text{ for } t = 1, 2, \ldots, \tau, \tag{1.12}$$

$$y_t = a_2' \hat{F}_t + b' W_t + \tilde{u}_t \ \text{ for } t = \tau + 1, \ldots, T \tag{1.13}$$

where $a_1' = a' D_1^{-}$ and $a_2' = a' D_2^{\dagger -}$, and $\tilde{u}_t = u_t + o_p(1)$.

If the number of factors is assumed to be known a priori, $\bar{r} = r$, then $D_1^{-} = D_1^{-1}$, $D_2^{\dagger -} = D_2^{\dagger -1}$. Since $D_1 \neq D_2^{\dagger}$, it follows that $D_1^{-1} \neq D_2^{\dagger -1}$ and thus $a_1 \neq a_2$. Therefore, using the indicator function $\mathbf{1}(t > \tau)$, (1.12) and (1.13) can be rewritten as

$$y_t = a_1' \hat{F}_t + (a_2 - a_1)' \hat{F}_t \mathbf{1}(t > \tau) + b' W_t + \tilde{u}_t, \ t = 1, 2, \ldots, T. \tag{1.14}$$

A straightforward implication of the previous result is that if we were to impose the number of factors, on a priori ground, therefore ignoring the set of regressors $\hat{F}_t\mathbf{1}(t > \tau)$ in (1.14), in general the estimation of b will become inconsistent due to omitted variables.

Remark 2. Interestingly, there are many examples in the literature where, for theoretical or practical reasons, the number of factors is imposed as a prior. For example, to name a few, a single common factor representing a global effect is assumed in the well-known study by Bernanke, Boivin and Eliasz (2005) on measuring the effects of monetary policy in Factor Augmented VAR (FAVAR) models, as well as in the risk analysis in portfolios of corporate debt by Gourieroux and Gagliardini (2011) where a single factor is supposed to capture a latent macro-variable. Likewise, two factors are a priori imposed by Rudebusch and Wu (2008) in their macro-finance model. Notice that a similar argument will render inconsistent the impulse response functions in FAVAR models where the regressand in (1.11) becomes $y_{t+1} = (F_{t+1}, W_{t+1})'$. □

Remark 3. Alternatively, if the number of factors is not imposed as a priori and instead is estimated by the IC of BN (2002), we will show in the next section (Proposition 1.2) that the estimated number of factors will tend to $r + k_1$ as the sample size grows. In this case, D_1 and $D_2^\dagger$ are $(r + k_1) \times r$ matrices, and by the definitions of D_1 and $D_2^\dagger$, it is easy to show that we can always find a $r \times (r + k_1)$ matrix $D^- = D_1^- = D_2^{\dagger -}$ such that $D^- D_1 = D^- D_2^\dagger = I_r$. If we define

$$\bar{a}' = a' D^-, \tag{1.15}$$

then $a_1' = a_2' = \bar{a}'$ so that (1.12) and (1.13) can be rewritten as

$$y_t = \bar{a}' \hat{F}_t + b' W_t + \tilde{u}_t, \ t = 1, 2, \ldots, T, \tag{1.16}$$

so that the estimation of (1.11) will not be affected by the estimated factors under big breaks if $\bar{r} = r + k_1$. □

Remark 4. Yet, even in this case, the factors themselves may be the direct subject of interest and thus their interpretation can have important implications for structural analysis. For example, a large body of empirical research in financial economics is concerned with identifying the factors that determine asset returns.[4] However, the existence of big breaks may hamper this identification procedure. For instance, if $\widehat{r} = 2$, a relevant question would be: are there two genuine factors or one factor and one break? Our testing procedure provides a useful tool to disentangle these two cases (see Section 4.4 below).

Another area where our testing approach could be useful is in applications where the estimated factors are modeled in a VAR in order to identify the structural shocks driving them

[4]Chen et al (1986) and Shanken and Weinstein (2006) are good illustrations of attempts to interpret the underlying forces in the stock market developments in terms of some observed macro variables.

(see e.g. Charnavoki and Dolado, 2012). This identification becomes more difficult as the number of factors increases. For this reason, it is very important to determine whether the selection of a large number of factors is due to having several genuine factors or to a break affecting some of them. The insight is that instead of having to identify $r + k_1$ shocks we would only have to identify r. □

Summing up, the use of estimated factors as the true factors when assuming that the number of factors is a priori known will lead to inconsistent estimates in a FAR under big breaks. As a simple remedy, $\hat{F}_t \mathbf{1}(t > \tau)$ should be added as regressors when big breaks are detected and the break date is located. Alternatively, without pretending to know a priori the true number of factors, the estimation of FAR will be robust to the estimation of factors under big breaks if the number of factors is chosen according to some consistent estimator. Yet, this may hinder the correct interpretation of the estimated factors in terms of observables. As a result, in order to run regression (1.16), our advice is to avoid imposing the number of factors a priori, unless a formal test of big breaks is implemented. We will illustrate these points in Section 1.5 by means of simulations in a typical forecasting exercise where the predictors are common factors estimated by PCA.

1.3.2 The estimated number of factors

BE (2011) have previously argued that the presence of structural breaks in the factor loadings may lead to the overestimation of the number of factors. Yet, since they do not provide a formal proof of this result, we proceed to fill this gap.

Let $\hat{r}$ be the estimated number of factors in (1.5) using the IC proposed by BN (2002), and ϑ^1_{NT} be the largest eigenvalue of $(NT)^{-1} \sum_{t=1}^{T} e_t e_t'$. Then, the following result holds:

Proposition 1.2. *Suppose* $\vartheta^1_{NT} = O_p(\delta^{-2}_{NT})$ *and Assumptions 1 to 5 hold, then:*

$$\lim_{N,T\to\infty} \mathbb{P}[\hat{r} = r + k_1] = 1.$$

Again, absent big breaks ($k_1 = 0$), this result trivially replicates Theorem 2 of BN (2002). However, under big breaks ($k_1 > 0$), their IC will overestimate the number of original factors by the number of big breaks ($0 < k_1 \leq r$) because, as shown above, a FM with this type of break admits a representation without a break but with more factors.

In sum, when we use PCA to estimate the factor space and the IC of BN (2002) to estimate the number of factors, the small breaks can be safely ignored, while the big breaks will lead to the inconsistencies of $\hat{F}_t$ and $\hat{r}$.

1.4 Testing for Structural Breaks

1.4.1 Hypotheses of interest and test statistics

Our goal here is to develop a test for big breaks. As mentioned above, if we were to follow the usual approach in the literature to test for structural breaks, we would consider the following null and alternative hypotheses in (1.1) and (1.2): $H_0 : A = B$ vs. $H_1 : A \neq B$. However, this standard formulation faces two problems. First, if only small breaks occur, the estimation and inference based on PAC are not affected. Thus, we can ignore these breaks. Secondly, and foremost, since A and B are $N \times r$ matrices, we would face an infinite-dimensional parameter problem as N grows if we were to consider differences in all their individual elements. To solve the first problem, we focus only on big breaks and consider $H_0 : k_1 = 0$ vs. $H_1 : k_1 > 0$, where the new null and alternative hypotheses correspond to the cases where there are no big breaks (yet there may be small breaks) and there is at least one big break, respectively.

Relying upon the discussion in Section 1.3.1 about the inconsistency of $\widehat{F}$ for the space of the true factors F when big breaks occur, our strategy to circumvent the second problem is to focus on how the dependence properties of the $\bar{r}$ estimated factors (using the whole sample) change before and after the potential break date. Since, in line with the standard assumption in FM, the number of true factors (r) is considered to be invariant to the sample size, our previous result in Proposition 1.2 ensures that $r + k_1$, with $k_1 \leq r$, is finite-dimensional.

To test the above null hypothesis, we consider the following two-step procedure:

1. In the first step, the number of factors to estimate, $\bar{r}$, is determined and $\bar{r}$ common factors ($\hat{F}_t$) are estimated by PCA.

2. In the second step, we consider the following linear regression of one of the estimated factors on the remaining $\bar{r}-1$ ones. For example, using the first factor as the regressand, this leads to the regression:

$$\hat{F}_{1t} = c_2\hat{F}_{2t} + \cdots + c_{\bar{r}}\hat{F}_{\bar{r}t} + u_t = c'\hat{F}_{-1t} + u_t \tag{1.17}$$

   where $\hat{F}_{-1t} = [\hat{F}_{2t}, \ldots, \hat{F}_{\bar{r}t}]'$ and $c = [c_2, \ldots, c_{\bar{r}}]'$ are $(\bar{r} - 1) \times 1$ vectors. Then we test for a structural break of c in the above regression. If a structural break is detected, then we reject $H_0 : k_1 = 0$; otherwise, we cannot reject the null of no big breaks.

Remark 5. In the first stage, we have recommended to choose $\bar{r}$ as some consistent estimator of r to obtain the best size and power properties. Although our Proposition 1.2 is based on the $\hat{r}$ estimated by the IC of BN (2002), one can also use other procedures to consistently estimate r. For example, Onatski (2009, 2010), Ahn and Horenstein (2013) show that their

methods have better finite sample properties than BN (2002) especially when the errors are cross sectionally correlated. That being said, as will be discussed below, one important feature of our tests is that they do not rely on the correct estimation of r. In the second step, although there are many methods of testing for breaks in a simple linear regression model, we follow Andrews (1993) to define the LM and Wald tests when the possible break date is assumed to be known, and their *Sup-type* versions when there is no prior knowledge about it. Moreover, since the LM, Wald and LR test statistics have the same asymptotic distribution under the null, we focus on the first two because they are computationally simpler. □

Define $D^* = V^{-1/2}\Upsilon'\Sigma_A^{1/2}$ as the limit of the matrix D in equation (1.7), where $V = diag(v_1, v_2, \ldots, v_r)$, $v_1 > v_2 > \ldots > v_r$ are the eigenvalues of $\Sigma_A^{1/2}\Sigma_F\Sigma_A^{1/2}$, and Υ is the corresponding eigenvector matrix (see Bai, 2003). Define $\mathcal{F}_{1t} = D_1^* F_t$ and $\mathcal{F}_{-1t} = D_{-1}^* F_t$, where D_1^* is the first row of D^* and D_{-1}^* is the matrix containing the second to last rows of D^*. Finally, let the $(r-1)\times(r-1)$ matrix $S = \lim \mathrm{Var}\big(\frac{1}{T}\sum_{t=1}^{T}\mathcal{F}_{-1t}\mathcal{F}_{1t}\big)$.

Following Andrews (1993), the LM test statistic is then defined as follows:

$$\mathcal{L}(\bar{\pi}) = \frac{1}{\bar{\pi}(1-\bar{\pi})}\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{\tau}\hat{F}_{-1t}\hat{u}_t\Big)'\hat{S}^{-1}\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{\tau}\hat{F}_{-1t}\hat{u}_t\Big) \tag{1.18}$$

where $\bar{\pi} = \tau/T$, τ is a pre-assumed date for the potential break, $\hat{u}_t$ is the residual in the OLS regression (1.17), which by construction equals $\hat{F}_{1t}$, and $\hat{S}$ is a consistent estimator of S.[5]

The corresponding Sup-LM statistic is defined as:

$$\mathcal{L}(\Pi) = \sup_{\pi\in\Pi}\frac{1}{\pi(1-\pi)}\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{[T\pi]}\hat{F}_{-1t}\hat{u}_t\Big)'\hat{S}^{-1}\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{[T\pi]}\hat{F}_{-1t}\hat{u}_t\Big) \tag{1.19}$$

where Π is any set whose closure lies in $(0,1)$.

Similarly, the Wald and Sup-Wald test statistics can be constructed as:

$$\mathcal{L}^*(\bar{\pi}) = \bar{\pi}(1-\bar{\pi})\cdot T\Big(\hat{c}_1(\bar{\pi}) - \hat{c}_2(\bar{\pi})\Big)'\hat{S}^{-1}\Big(\hat{c}_1(\bar{\pi}) - \hat{c}_2(\bar{\pi})\Big) \tag{1.20}$$

and

$$\mathcal{L}^*(\Pi) = \sup_{\pi\in\Pi}\pi(1-\pi)\cdot T\Big(\hat{c}_1(\pi) - \hat{c}_2(\pi)\Big)'\hat{S}^{-1}\Big(\hat{c}_1(\pi) - \hat{c}_2(\pi)\Big) \tag{1.21}$$

where $\hat{c}_1(\pi)$ and $\hat{c}_2(\pi)$ are OLS estimates of c using subsamples before and after the break point : $[T\pi]$.[6]

[5] See Appendix A.3 for discussions on the estimation of S.

[6] We can also construct the Wald test as $T\Big(\hat{c}_1(\bar{\pi}) - \hat{c}_2(\bar{\pi})\Big)'\Big(\frac{\hat{S}_1}{\bar{\pi}} + \frac{\hat{S}_2}{(1-\bar{\pi})}\Big)^{-1}\Big(\hat{c}_1(\bar{\pi}) - \hat{c}_2(\bar{\pi})\Big)$ and the Sup-Wald test similarly, where $\hat{S}_1$ and $\hat{S}_2$ are estimates of S using subsamples. Yet, in all our simulations, the results based on these two methods are very similar. Therefore, for brevity, we focus on the ones obtained using the full sample estimation of S, as in (1.20) and (1.21).

To illustrate why our two-step testing procedure is able to detect the big breaks, it is useful to consider a simple example where $r = 1$, $k_1 = 1$ (one common factor and one big break). Then (1.5) becomes:

$$X_t = Af_t + Bg_t + e_t$$

where $g_t = 0$ for $t = 1, \ldots, \tau$, and $g_t = f_t$ for $t = \tau + 1, \ldots, T$. By Proposition 1.2, we will tend to get $\hat{r} = 2$ in this case. Suppose now that we estimate 2 factors ($\bar{r} = 2$). Then, by Proposition 1.1, we have:

$$\begin{pmatrix} \hat{f}_{1t} \\ \hat{f}_{2t} \end{pmatrix} = D \begin{pmatrix} f_t \\ g_t \end{pmatrix} + o_p(1)$$

where $D = \begin{pmatrix} d_1 & d_2 \\ d_3 & d_4 \end{pmatrix}$ is a non-singular matrix. By the definition of g_t we have:

$$\hat{f}_{1t} = d_1 f_t + o_p(1) \quad \hat{f}_{2t} = d_3 f_t + o_p(1) \quad \text{for} \quad t = 1, \ldots, \tau,$$

$$\hat{f}_{1t} = (d_1 + d_2) f_t + o_p(1) \quad \hat{f}_{2t} = (d_3 + d_4) f_t + o_p(1) \quad \text{for} \quad t = \tau + 1, \ldots, T,$$

which imply that:

$$\hat{f}_{1t} = \frac{d_1}{d_3} \hat{f}_{2t} + o_p(1) \quad \text{for} \quad t = 1, \ldots, \tau,$$

$$\hat{f}_{1t} = \frac{d_1 + d_2}{d_3 + d_4} \hat{f}_{2t} + o_p(1) \quad \text{for} \quad t = \tau + 1, \ldots, T.$$

Thus, we can observe that the two estimated factors are linearly related and that the coefficients $\frac{d_1}{d_3}$ and $\frac{d_1+d_2}{d_3+d_4}$ before and after the break date must be different due to the non-singularity of the D matrix. As a result, regressing one of the estimated factors on the other and testing for a structural break in this regression, we should reject the null of no big break. In the case where $d_3 = 0$, the above argument fails. But since d_1 and d_4 are non-zeros (otherwise D will have reduced-rank), the estimated slope in the first subsample will diverge while it will converge to some bounded number in the second subsample. Therefore our test also has power in this case.[7]

Likewise, if the break date τ is not a priori assumed to be known, the Sup-type tests will yield a natural estimate of τ at the date when the test reaches its maximum value. In what follows, we derive the asymptotic distribution of the test statistics (1.18) to (1.21) under the null hypothesis, as well as extend the intuition behind this simple example to the more general case to show that our tests have power against relevant alternatives.

1.4.2 Limiting distributions under the null hypothesis

Since in most applications the number of factors is estimated by means of BN's (2002) IC, and it converges to the true one under the null hypothesis of no big break, we start with the

[7] This is the case for the Wald test but may not be true for the LM test, because our Wald test is directly built upon the difference between the estimated coefficients.

most interesting case where $\bar{r} = r$. To derive the asymptotic distributions of the LM and Wald statistics, we adopt the following additional assumptions:

Assumption 6. $\sqrt{T}/N \to 0$ *as* $N \to \infty$ *and* $T \to \infty$.

Assumption 7. $\{F_t\}$ *is a stationary and ergodic sequence, and* $\{F_{it}F_{jt} - E(F_{it}F_{jt}), \Omega_t\}$ *is an adapted mixingale with* γ_m *of size* -1 *for* $i, j = 1, 2, \ldots, r$, *that is:*

$$\sqrt{E\Big(E(Y_{ij,t}|\Omega_{t-m})^2\Big)} \leq c_t\gamma_m$$

where $Y_{ij,t} = F_{it}F_{jt} - E(F_{it}F_{jt})$, Ω_t *is a* $\sigma-$*algebra generated by the information at time* $t, t-1, \ldots,$ $\{c_t\}$ *and* $\{\gamma_m\}$ *are non-negative sequences and* $\gamma_m = O(m^{-1-\delta})$ *for some* $\delta > 0$.

Assumption 8. $\sup_{\pi\in[0,1]} \left\| \frac{1}{\sqrt{NT}} \sum_{t=1}^{T\pi} \sum_{i=1}^{N} \alpha_i F_t' e_{it} \right\|^2 = O_p(1)$.

Assumption 9. $\left\|\hat{S} - S\right\| = o_p(1)$, *and* S *is a* $(r-1) \times (r-1)$ *symmetric positive definite matrix.*

Assumption 10. *The eigenvalues of the* $r \times r$ *matrix* $\Sigma_A\Sigma_F$ *are distinct.*

Assumption 6 and 8 are required to bound the estimation errors of $\hat{F}_t$, while Assumption 7 is needed to derive the weak convergence of the test statistics using the Functional Central Limit Theorem (FCLT). Assumption 10 corresponds to Assumption G of Bai (2003), which is required for $D \xrightarrow{p} D^*$.

Note that these assumptions are not restrictive. Assumption 6 allows T to be $O(N^{1+\delta})$ for $-1 < \delta < 1$. As for Assumption 7, it allows us to consider a quite general class of linear processes for the factors: $F_{it} = \sum_{k=1}^{\infty} \theta_{ik} v_{i,t-k}$, where $v_t = [v_{1t} \ldots v_{rt}]'$ are i.i.d with zero means, and $Var(v_{it}) = \sigma_i^2 < \infty$. In this case, it can be shown that:

$$\sqrt{E\Big(E(Y_{ij,t}|\Omega_{t-m})^2\Big)} \leq \sigma_i\sigma_j \left(\sum_{k=m}^{\infty} |\theta_{ik}| \right) \left(\sum_{k=m}^{\infty} |\theta_{jk}| \right)$$

for which it suffices that

$$\left(\sum_{k=m}^{\infty} |\theta_{ik}| \right) = O(m^{-1/2-\delta})$$

for some $\delta > 0$, which is satisfied for a large class of ARMA processes. Assumption 8 is similar to Assumption F.2 of Bai (2003), which involves zero-mean random variables. Finally, a consistent estimate of S can be calculated by a HAC estimator such as Newey and West's (1987) estimator with a Barlett kernel, which is the one used in our simulations below.[8]

Let " $\xrightarrow{d}$ "denote *convergence in distribution*, then:

[8]Though not reported, other estimators, like those based on Parzen kernels, yield similar results in our simulations about the size and power properties of the LM and Wald tests.

Theorem 1.3. *Under the null hypothesis* $H_0 : k_1 = 0$ *and Assumptions 1 to 10, as* $N, T \to \infty$*, we have that both the LM and Wald tests verify*

$$\mathcal{L}(\bar{\pi}),\ \mathcal{L}^*(\bar{\pi}) \xrightarrow{d} \chi^2(r-1)$$

where $\bar{\pi} = \tau/T$ *for a given* τ*; and*

$$\mathcal{L}(\Pi),\ \mathcal{L}^*(\Pi) \xrightarrow{d} \sup_{\pi \in \Pi} \Big(\mathcal{W}_{r-1}(\pi) - \pi \mathcal{W}_{r-1}(1)\Big)' \Big(\mathcal{W}_{r-1}(\pi) - \pi \mathcal{W}_{r-1}(1)\Big)/[\pi(1-\pi)]$$

for any set Π *whose closure lies in* $(0,1)$*, where* $\mathcal{W}_{r-1}(\cdot)$ *is a* $r-1$ *vector of independent Brownian Motions on* $[0,1]$ *restricted to* Π*.*

The critical values for the Sup-type test are provided in Andrews (1993).

Remark 6. It is easy to show that Theorem 1.3 still holds when $\bar{r} < r$. However, when $\bar{r} > r$, the properties of $\hat{F}_{\bar{r}t}$ are unknown since the $\bar{r}$th eigenvectors of XX' may be related to the properties of e_t. Thus, the asymptotic distribution cannot be derived in a similar way. Yet, as the simulations in Section 1.5 show, in such an instance Theorem 1.3 still provides a reasonably good approximation for the distributions of our test statistics in finite samples. Moreover, the case where $\bar{r} > r$ can be avoided if, instead of relying on a priori choice of $\bar{r}$, practitioners use BN's (2002) IC or other consistent estimators of r, in line with Proposition 1.2. □

1.4.3 Performance of the tests under the alternative hypothesis

We now extend the insight of the simple example considered in Section 1.4.1 to show that, under the alternative hypothesis ($k_1 > 0$), the linear relationship between the estimated factors changes at time τ, so that the proposed tests are able to detect big breaks.

Assuming that $r < \bar{r} \leq r + k_1$, then the matrix D_1 and $D_2^\dagger$ in (1.9) and (1.10) become $\bar{r} \times r$ matrices. Notice that since $\bar{r} > r$ we can always find $\bar{r} \times 1$ vectors ρ_1 and ρ_2 which belong to the null spaces of D_1' and $D_2^{\dagger\prime}$ separately, that is, $\rho_1' D_1 = 0$ and $\rho_2' D_2^\dagger = 0$. Hence, premultiplying both sides of (1.9) and (1.10) by ρ_1' and ρ_2' leads to:

$$\begin{aligned} \rho_1' \hat{F}_t &= o_p(1) \quad t = 1, 2, \ldots, \tau, \\ \rho_2' \hat{F}_t &= o_p(1) \quad t = \tau + 1, \ldots, T \end{aligned}$$

which, after normalizing one of the elements of ρ_1 and ρ_2 (e.g., the first one) to be 1, implies that:

$$\hat{F}_{1t} = \hat{F}'_{-1t} \rho_1^* + o_p(1) \quad t = 1, 2, \ldots, \tau, \tag{1.22}$$

$$\hat{F}_{1t} = \hat{F}'_{-1t} \rho_2^* + o_p(1) \quad t = \tau + 1, \ldots, T, \tag{1.23}$$

where $\hat{F}'_{-1t} = [\hat{F}_{2t}, \ldots, \hat{F}_{\bar{r}t}]$ and ρ_1^*, ρ_2^* are both $(\bar{r}-1) \times 1$ vectors. Next, to show that $\rho_1^* \neq \rho_2^*$, we proceed as follows. Suppose that $\gamma \in \text{Null}(D_1')$ and $\gamma \in \text{Null}(D_2^{\dagger\prime})$, then by the definitions of D_1 and $D_2^\dagger$ and by the basic properties of full-rank matrices, it holds that $\gamma \in \text{Null}(D')$. Since D is a full rank $\bar{r} \times (r + k_1)$ matrix and $\bar{r} \leq r + k_1$, then $\text{Null}(D') = 0$ and thus $\gamma = 0$. Therefore, the only vector that belongs to the null space of D_1 and $D_2^\dagger$ is the trivial zero vector. Further, because the rank of the null space of D_1 and $D_2^\dagger$ is larger than 1, we can always find two non-zero vectors such that $\rho_1 \neq b\rho_2$ for any constant $b \neq 0$.

Notice that when $\bar{r} \leq r$, the rank of the null spaces of D_1 and $D_2^\dagger$ will possibly become zero. Hence, the preceding analysis does not apply in this case despite the existence of linear relationships among the estimated factors. If we regress one of the estimated factors on the others, with $\hat{\rho}_1$ and $\hat{\rho}_2$ denoting the OLS estimates of the coefficients using the two subsamples before and after the break, in general we cannot verify that $\text{plim}\hat{\rho}_1 \neq \text{plim}\hat{\rho}_2$.

Remark 7. One underlying assumption in the above argument is that one of the elements of ρ_1 and ρ_2 (e.g., the first ones) are different from zero. This assumption is hard to verify since the D matrix depends on Γ and F^* in a highly nonlinear way, [9] and it is not difficult to find DGPs where this assumption does not hold. This normalization issue makes it really hard to come up with a formal result on the consistency of our test. Instead, we have run a large number of simulations to study the actual power properties of our tests for various DGPs, including the ones where the first elements of ρ_1 and ρ_2 are both zeros. The general finding is that our Wald and Sup-Wald tests are very powerful for all the DGPs we have considered, while the LM and Sup-LM tests may lose powers when the normalization issue arises. We present some of the representative simulation results in Section 1.5 but more results are available upon request. □

Remark 8. Since our tests are based on a linear regression model, many other available methods in the literature can also be applied in our second-stage procedure. For instance, when the break date is not known as a priori, one can use the CUSUM type-test first proposed by Brown, Durbin and Evans (1975), and also Chen and Hong's (2012) test via nonparametric regression. Thus, this flexibility allows practitioners to draw conclusions about breaks based on broader evidence than that just provided by a single test. □

Remark 9. Although our tests have been designed for a single common break at date τ, they should also exhibit powers against other interesting alternatives such as multiple breaks and a change in the number of factors.[10] □

[9] By the result of Bai (2003), when $\bar{r} = r + k_1$, the $(r + k_1) \times (r + k_1)$ matrix $D = V_{NT}^{-1}(\hat{F}^{*\prime} F^*/T)(\Gamma'\Gamma/N)$, where V_{NT} is the diagonal matrix with the first $r + k_1$ eigenvalues of $X'X/NT$.

[10] See our online appendix for details.

1.4.4 Disentangling breaks in loadings from breaks in factors

A potential critique for all available tests of big breaks in FM is that they cannot differentiate between breaks in factor loadings and breaks in the covariance matrix of factors. For illustrative purposes, let us consider a FM with $r = 2$, where the factor loadings are constant but the covariance matrix of the factors breaks, such that: $E(F_t F_t') = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ for $t = 1, \ldots, T/2$, and $E(F_t F_t') = \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$ for $t = T/2 + 1, \ldots, T$, with $\rho \notin \{0, 1\}$. If we further assume $\Sigma_A = \lim_{N\to\infty} A'A/N$ is a diagonal matrix, then, in view of Bai (2003), we have that $\hat{F}_t = F_t + o_p(1)$, where $\hat{F}_t$ is a 2×1 vector. In this case, regressing $\hat{F}_{1t}$ on $\hat{F}_{2t}$ will yield estimated coefficients close to ρ and $-\rho$ before and after the break, respectively. As a result, our tests will reject the null of no big break in the loadings while the true DGP has a break in the factors.

Although the above example has been excluded by our Assumptions 2 and 7, it could well be that the factor dynamics are subject to structural breaks in practice. For instance, if the factors are interpreted as common shocks to the economy, then it is likely that their volatilities may have decreased since the beginning of 1980s, as evidenced by recent studies on the effect of the Great Moderation (see, e.g., Gali and Gambetti, 2009). Hence, for interpretational purposes, it becomes relevant to identify which is the source of breaks.

Assuming that there is only one source of instability and that one break in the FM has been detected at some date τ, one can differentiate between the break in the loadings and the break in the factor dynamics by comparing the number of factors obtained with the whole sample to those in each of the subamples split by τ. To see this, notice that, absent a big break in the factor loadings, the number of factors will be consistently estimated for the whole sample and each sub samples, as long as the factors satisfy:

$$\frac{1}{\tau}\sum_{t=1}^{\tau} F_t F_t' \xrightarrow{p} \Sigma_F^1 > 0, \text{ and } \frac{1}{T-\tau}\sum_{t=\tau+1}^{T} F_t F_t' \xrightarrow{p} \Sigma_F^2 > 0.$$

One important observation is that, when the factors are stationary and the rejection of the null is due to big breaks in the loadings, the true number of factors, r, can be consistently estimated for each of the two subsamples while, for the whole sample, it will be overestimated in light of Proposition 1.2. Therefore, for a given data set, if $\hat{r} = 2$ in the whole sample and our test rejects the null with $\bar{r} = 2$, it could be that there is one factor and one big break, or two factors with changing correlation, as shown above. In the first case, the estimated numbers of factors in each of the two subsamples converge to 1, while for the latter case it converges to 2. Thus, these different rank properties could become the basis of our identification strategy for the source of the break.

When both breaks in the loadings and factors exist, our strategy of considering a finite dimensional linear regression still works, but the asymptotic distribution of the proposed test statistics may be different. The reason is that the estimated factors, which are consistent estimators of the true factor space under the null, may also experience breaks in their dynamics. As a result, Hansen's (2000) results imply that the asymptotic distributions of the Wald and Sup Wald tests may change if the dynamics of regressors (the estimated factors in our case) have breaks. We leave this interesting question for future research.

1.5 Simulations

In this section, we first use a simple factor-based forecasting model to illustrate the consequences of ignoring big breaks, as discussed in Section 3. Next, we study the finite sample properties of our proposed LM/Wald and Sup-LM/Wald tests. We pay special attention to the the sizes and the powers when $\bar{r} > r$ since, as discussed previously, obtaining formal results for this case is very difficult. A comparison with the tests of BE (2011) and HI (2012) is also provided to illustrate the advantages of our tests in term of power. Throughout this section, the potential breaking date is considered to located at half of the sample ($\tau = T/2$) and is taken to be a priori known for the LM/Wald tests while Π is chosen as $[0.15, 0.85]$ for the Sup-type versions. Finally, the covariance matrix S is estimated using the HAC estimator of Newey and West (1987).

1.5.1 The effect of big breaks on forecasting

In this section we consider the effect of having big breaks in a typical forecasting exercise where the predictors are estimated common factors. First, we have a large panel of data X_t driven by the factors F_t which are subject to a big break in the factor loadings:

$$X_t = AF_t\mathbf{1}(t \leq \tau) + BF_t\mathbf{1}(t > \tau) + e_t.$$

Secondly, the variable we wish to forecast y_t, which is excluded from to X_t, is assumed to be related to F_t as follows:

$$y_{t+1} = a'F_t + v_{t+1}.$$

We consider a DGP where $N = 100$, $T = 200$, $\tau = 100$, $r = 2$, $a' = [1 \quad 1]$, F_t are generated as two AR(1) processes with coefficients 0.8 and 0.2, respectively, e_t and v_t are i.i.d standard normal variables, and the break size is characterized by a range of mean shifts between 0 and 1.

The following forecasting procedures are compared in our simulation:

Benchmark Forecasting: The factors F_t are treated as observed and are used directly as predictors. The one-step-ahead forecast of y_t is defined as $\hat{y}_{t+1|t} = \hat{a}'F_t$, where $\hat{a}$ is the OLS estimate of a in the regression of y_{t+1} on F_t.

Forecasting 1: We first estimate 2 factors $\hat{F}_t$ from X_t by PCA, which are then used as predictors in $\hat{y}_{t+1|t} = \hat{a}'\hat{F}_t$, where $\hat{a}$ is the OLS estimate of a in the regression of y_{t+1} on $\hat{F}_t$.

Forecasting 2: We first estimate 2 factors $\hat{F}_t$ from X_t by PCA, and then use $\hat{F}_t$ and $\hat{F}_t\mathbf{1}(t > \tau)$ as predictors. $\hat{y}_{t+1|t} = \hat{a}'[\hat{F}_t' \quad \hat{F}_t\mathbf{1}(t > \tau)']'$, where $\hat{a}$ is the OLS estimate of a in the regression of y_{t+1} on $\hat{F}_t$ and $\hat{F}_t\mathbf{1}(t > \tau)$.

Forecasting 3: We first estimate 4 factors (replicating $r + k_1 = 4$) $\hat{F}_t$ from X_t by PCA, which are then used as predictors in $\hat{y}_{t+1|t} = \hat{a}'\hat{F}_t$, where $\hat{a}$ is the OLS estimate of a in the regression of y_{t+1} on $\hat{F}_t$.

The above forecasting exercises are implemented recursively, e.g., at each time t, the data $X_t, X_{t-1}, \ldots, X_1$ and $y_t, y_{t-1}, \ldots, y_1$ are treated as known to forecast y_{t+1}. This process starts from $t = 149$ to $t = 199$, and the mean square errors (MSEs) are calculated by

$$MSE = \sum_{t=149}^{199} \frac{(y_{t+1} - \hat{y}_{t+1|t})^2}{51}.$$

To facilitate the comparisons, the MSEs of the Benchmark Forecasting is standardized to be 1.

The results obtained from 1000 replications are reported in Figure 1.1, plotting MSEs against the different break sizes in the above-mentioned range. It is clear that the MSEs of the Forecasting 1 method increases drastically with the size of the breaks, in line with our discussion in Section 1.3. By contrast, the Forecasting 2 and 3 procedures perform equally well and their MSEs remain constant as the break size increases. Notice, however, that they cannot outperform the benchmark forecasting due to the estimation errors of the factors for the chosen sizes of N and T. In line with our previous analysis, the lesson to be drawn from this exercise is that, in case of a big break, imposing the number of factors a priori can significantly worsen forecasts.

1.5.2 Size properties

We first simulate data from the following DGP:

$$X_{it} = \sum_{k=1}^{r} \alpha_{ik} F_{kt} + e_{it}$$

TABLE 1.1: Empirical Sizes of the Nominal 5% Size Tests for Different Choices of $\overline{r}$ when $r = 3$.

N	T	$\hat{\alpha}_{0.05}\|\bar{r}=2$				$\hat{\alpha}_{0.05}\|\bar{r}=3$				$\hat{\alpha}_{0.05}\|\bar{r}=4$			
		LM	Sup LM	Wald	Sup Wald	LM	Sup LM	Wald	Sup Wald	LM	Sup LM	Wald	Sup Wald
100	100	5.0	1.0	5.9	4.8	2.3	0.2	4.2	6.7	0.5	0.2	1.3	11.6
100	150	5.0	1.9	4.9	3.1	3.5	0.7	3.7	4.8	1.1	0.3	1.9	7.0
100	200	5.7	2.7	5.0	4.0	4.9	1.8	4.0	3.5	3.0	0.5	2.9	3.9
100	250	5.3	3.2	5.3	3.9	4.4	1.8	4.7	3.2	2.3	0.9	3.4	3.1
100	300	6.2	4.5	6.7	4.0	5.3	2.0	5.1	3.4	3.8	1.1	4.7	3.9
150	100	5.3	1.2	5.9	5.1	2.6	0.2	4.0	7.9	0.8	0.2	2.3	12.9
150	150	5.9	1.8	5.2	4.0	2.9	0.5	3.4	4.0	1.3	0.3	2.7	6.1
150	200	5.5	2.6	6.2	4.5	3.5	1.2	5.1	3.4	2.3	0.9	3.0	4.3
150	250	6.0	2.9	6.9	3.8	3.5	1.6	5.7	3.1	3.2	0.5	3.6	4.7
150	300	5.8	3.7	6.3	4.4	3.9	2.5	5.1	4.0	3.5	1.3	4.0	3.7
200	100	4.6	1.1	5.4	5.0	2.3	0.1	3.0	8.6	0.4	0.4	1.5	15.6
200	150	4.7	2.3	5.6	3.2	2.8	0.2	3.7	4.3	1.2	0.1	2.7	5.6
200	200	5.4	3.0	5.1	2.9	4.0	1.6	3.4	2.5	2.6	1.3	3.2	3.5
200	250	6.2	3.7	7.0	4.0	3.8	2.0	6.8	4.1	2.4	1.1	4.1	5.2
200	300	5.3	3.1	5.5	4.6	3.2	1.5	3.5	4.0	3.4	1.3	2.6	4.5
250	100	5.2	0.8	7.4	5.1	2.1	0.4	4.5	7.0	0.6	0.2	3.5	12.9
250	150	4.1	2.5	5.7	3.6	2.9	0.5	3.9	4.2	1.6	0.0	2.4	6.4
250	200	5.3	2.6	6.5	4.9	3.5	0.8	4.6	5.0	2.9	0.3	3.4	5.2
250	250	5.3	3.1	6.2	4.3	4.7	1.8	5.6	3.1	4.0	0.7	3.5	3.6
250	300	5.5	4.0	5.1	3.7	4.3	1.5	4.0	3.3	3.4	1.4	2.9	3.7
300	100	4.7	0.6	5.2	5.4	1.5	0.2	3.4	8.5	0.3	0.3	2.9	14.0
300	150	4.6	1.8	6.4	5.4	2.9	0.8	4.8	4.7	1.7	0.5	2.8	7.0
300	200	3.7	2.6	7.0	4.0	3.2	0.8	6.5	4.1	1.7	0.5	4.2	5.5
300	250	5.9	3.5	6.3	4.1	4.8	1.7	5.2	3.4	2.7	1.0	3.3	3.5
300	300	5.7	4.2	4.2	4.1	6.2	3.2	4.4	3.4	3.9	1.4	2.8	3.2
1000	1000	5.7	6.1	7.1	5.9	5.8	4.2	6.2	4.9	6.5	4.7	5.8	3.5

Notes: The DGP is $X_{it} = \sum_{k=1}^{3} \alpha_{ik} F_{kt} + e_{it}$ where $F_{kt} = \phi_k F_{k,t-1} + v_{kt}$, $\alpha_{ik}, e_{it}, v_{kt} \sim i.i.d\ N(0,1)$, and $[\phi_1, \phi_2, \phi_3] = [0.8, 0.5, 0.2]$ (See Section 1.5.2).

where $r = 3$, α_{ik} and e_{it} are generated as i.i.d standard normal variables, and $\{F_{kt}\}$ are generated as:

$$F_{kt} = \phi_k F_{k,t-1} + v_{kt}$$

where $[\phi_1, \phi_2, \phi_3] = [0.8, 0.5, 0.2]$, and v_{kt} is another i.i.d standard normal error term. The number of replications is 1000. We consider both the LM and Wald tests and their Sup-type versions.

Table 1.1 reports the empirical sizes (in percentages) for the LM/Wald tests and Sup-LM/Wald tests using 5% asymptotic critical values for sample sizes (N and T) equal to 100, 150, 200, 250, 300 and 1000.[11] We consider three cases regarding the choice of the number of factors to be estimated by PC: (i) the correct one ($\bar{r} = r = 3$), (ii) smaller than the true number of factors ($\bar{r} = 2 < r = 3$), and (iii) larger than the true number of factors ($\bar{r} = 4 > r = 3$).[12] Although, for $N, T \geq 100$, BN's (2002) IC very often select 3 factors, there are some cases where 2 and 4 are also selected.

[11] As mentioned earlier, the critical values of the Sup-type tests are taken from Andrews (1993).

[12] Notice that the choice of $r = 3$ allows us to analyze the consequences of performing our proposed test with the under-parameterized choice of $\overline{r} = 2$, where two factors are needed to perform the LM/Wald tests. Had we chosen $r = 2$ as the true number of factors, then the test could not be implemented for $\bar{r} = 1$.

Broadly speaking, the LM and Wald tests are slightly undersized for $r = 2$ and 3, and especially so when $r = 4$. Yet, the empirical sizes converge to the nominal size as N and T increase. This finite sample problem is more accurate with the Sup-LM test especially for small T, in line with the findings in other studies (see, Diebold and Chen, 1996). This is hardly surprising because, for instance, when $T = 100$ and $\Pi = [0.15, 0.85]$, we only have 15 observations in the first subsample. By contrast, although the Sup-Wald test is too liberal for $T = 100$, in general it behaves better than the Sup-LM test (see Kim and Perron, 2009). Theoretically, the asymptotic distribution in Theorem 1.3 applies only for $\bar{r} = 2, 3$. Yet, the results in Table 1.1 show that this distribution also provides reasonably good approximations for the case $\bar{r} = 4$, the only exception being the Sup-Wald test, which is oversized for $T = 100$.

To further study how the size of our Wald tests is affected by the selection of $\bar{r}$ in a more general setup, we repeat the above simulations by generating the idiosyncratic errors in the following way as in BN (2002):

$$e_{it} = \rho e_{i,t-1} + u_{it} + \beta \sum_{h=i-J}^{i-1} u_{ht} + \beta \sum_{h=i+1}^{i+J} u_{ht}$$

where $u_{it} \sim i.i.d\ N(0, 1)$, β and J control the cross sectional correlation of e_{it}, and ρ controls the serial correlation of e_{it}. The other parts of the DGP remain the same as above, and we consider there different types of e_t: (1) only serially correlated ($\rho = 0.2, 0.5, 0.7$), (2) only cross sectionally correlated ($\beta = 0.2, 0.5, J = 5, 10$), (3) both cross sectionally and serially correlated. To save space, we fix $N = 100$ (similar to the data set in our application) and focus on the role of T.[13]

[13] More simulation results with different N and T are available upon request.

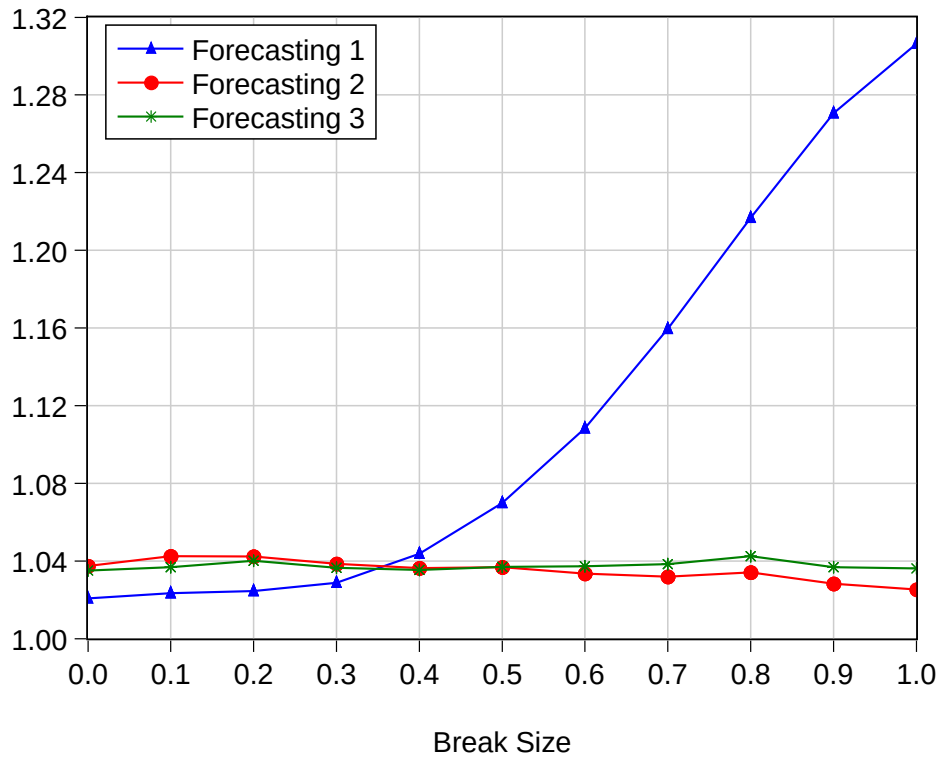


FIGURE 1.1: The MSEs of different forecasting methods in the presence of big breaks (see Section 1.5.1).

TABLE 1.2: Empirical Size of the Nominal 5% Size Tests for Different Choices of $\bar{r}$ when $r = 3$, and when the idiosyncratic errors are cross sectionally and serially correlated.

	$N = 100, T = 50$						$N = 100, T = 100$						$N = 100, T = 200$					
	Wald			Sup Wald			Wald			Sup Wald			Wald			Sup Wald		
	$\bar{r}=2$	$\bar{r}=3$	$\bar{r}=4$	$\bar{r}=2$	$\bar{r}=3$	$\bar{r}=4$	$\bar{r}=2$	$\bar{r}=3$	$\bar{r}=4$	$\bar{r}=2$	$\bar{r}=3$	$\bar{r}=4$	$\bar{r}=2$	$\bar{r}=3$	$\bar{r}=4$	$\bar{r}=2$	$\bar{r}=3$	$\bar{r}=4$
$\rho = 0.2, \beta = J = 0$	5.9	3.8	6.3	11.4	25.0	45.2	6.0	2.9	2.2	4.3	8.2	13.8	5.3	4.0	4.0	3.7	3.0	4.0
$\rho = 0.5, \beta = J = 0$	6.9	3.6	13.2	11.7	24.2	54.1	5.6	3.2	4.0	4.7	8.4	21.6	5.1	4.1	4.4	4.2	3.5	6.2
$\rho = 0.7, \beta = J = 0$	8.6	7.0	30.2	12.7	26.2	63.7	6.0	3.1	11.0	4.7	7.8	37.7	5.1	3.7	7.6	3.7	3.8	14.0
$\rho = 0, \beta = 0.2, J = 5$	5.9	4.6	4.0	12.8	22.9	46.2	5.5	2.8	1.4	4.7	6.9	13.5	4.9	3.6	3.9	3.7	3.4	3.5
$\rho = 0, \beta = 0.2, J = 10$	6.1	3.7	4.3	12.3	26.1	48.3	5.2	2.3	1.4	5.9	8.4	13.8	4.8	3.6	2.9	3.6	3.8	5.0
$\rho = 0, \beta = 0.5, J = 10$	5.2	3.4	4.1	11.4	26.7	52.7	5.7	3.2	2.2	4.9	6.8	14.5	6.2	4.5	4.0	4.6	3.4	4.7
$\rho = 0.2, \beta = 0.2, J = 5$	6.3	4.6	4.7	12.1	23.3	44.2	5.4	2.9	1.4	4.7	6.9	12.9	5.1	3.7	3.6	4.1	3.8	4.5
$\rho = 0.2, \beta = 0.5, J = 10$	6.2	3.3	4.1	11.1	26.7	51.6	6.1	4.2	2.3	4.9	7.8	15.7	5.9	4.8	4.2	3.6	3.5	3.5
$\rho = 0.5, \beta = 0.2, J = 5$	7.6	4.4	7.6	12.7	24.7	46.0	5.7	3.0	2.7	5.3	7.6	15.5	5.2	3.8	4.2	3.5	3.6	5.8
$\rho = 0.5, \beta = 0.5, J = 10$	6.2	4.8	7.9	14.7	29.2	51.7	7.2	4.6	3.9	5.8	9.8	16.5	6.1	5.0	4.1	4.7	4.8	6.4

Notes: The DGP is $X_{it} = \sum_{k=1}^{3} \alpha_{ik} F_{kt} + e_{it}$ where $F_{kt} = \phi_k F_{k,t-1} + v_{kt}$, and $[\phi_1, \phi_2, \phi_3] = [0.8, 0.5, 0.2]$. $e_{it} = \rho e_{i,t-1} + u_{it} + \beta \sum_{h=i-J}^{i-1} u_{ht} + \beta \sum_{h=i+1}^{i+J} u_{ht}$, where $\alpha_{ik}, u_{it}, v_{kt} \sim i.i.d\ N(0,1)$ (See Section 1.5.2).

TABLE 1.3: Empirical Power of Nominal 5% Size Tests for Different Choices of $\bar{r}$ when $r = 2$ and $k_1 = 2$.

N	T	$\hat{\alpha}_{0.05}\|\bar{r}=2$				$\hat{\alpha}_{0.05}\|\bar{r}=3$				$\hat{\alpha}_{0.05}\|\bar{r}=4$			
		LM	Sup LM	Wald	Sup Wald	LM	Sup LM	Wald	Sup Wald	LM	Sup LM	Wald	Sup Wald
100	100	6.3	1.8	8.1	5.4	77.9	1.8	100	98.3	41.7	0.5	100	97.3
100	150	8.9	2.5	10.0	4.8	95.8	24.0	100	100	88.8	2.8	100	99.9
100	200	8.9	4.1	9.3	5.4	97.6	72.9	92.0	92.0	95.5	39.6	91.8	92.5
100	250	12.0	5.3	12.4	6.5	99.1	98.0	97.4	97.4	99.0	77.9	97.4	97.4
100	300	13.0	6.5	11.6	6.0	99.6	98.0	83.6	83.6	99.4	94.1	83.5	83.7
150	100	6.1	2.2	7.8	5.9	77.9	1.4	99.7	99.5	41.6	0.6	99.8	99.0
150	150	7.5	2.2	8.3	5.0	95.4	24.5	100	100	88.5	2.2	100	100
150	200	8.8	4.1	9.8	5.4	98.8	76.5	100	100	97.7	40.2	100	100
150	250	9.7	4.8	10.3	6.0	99.4	94.4	99.0	99.1	98.5	79.1	99.0	99.1
150	300	11.4	6.3	10.8	7.1	99.7	98.6	90.5	91.1	99.7	94.5	90.7	91.1
200	100	6.4	1.5	7.6	4.6	79.4	2.3	100	97.7	42.9	0.7	100	99.2
200	150	8.5	3.4	9.5	6.3	97.0	24.1	100	100	89.0	3.0	100	100
200	200	8.6	3.5	9.3	4.5	99.0	77.6	100	100	98.0	38.8	100	100
200	250	11.5	4.5	12.3	5.7	100	96.8	100	100	100	82.7	100	100
200	300	11.2	5.4	12.6	6.4	99.8	98.8	99.9	99.9	99.7	95.1	99.9	99.9
250	100	5.1	1.4	6.7	4.5	80.4	1.8	100	99.7	45.2	1.0	100	99.2
250	150	6.7	2.4	7.8	5.0	97.0	24.5	99.9	100	90.7	3.2	100	100
250	200	7.2	3.4	7.8	5.0	99.2	78.9	100	100	98.4	40.9	100	100
250	250	10.5	5.5	11.3	5.8	99.8	95.6	100	100	99.7	82.4	100	100
250	300	11.5	5.7	12.0	7.6	99.9	99.2	100	100	99.9	95.1	100	100
300	100	6.0	1.6	7.0	6.7	80.1	1.2	100	99.1	45.4	0.3	100	98.9
300	150	8.6	2.1	9.9	4.7	97.3	24.9	100	100	91.5	3.4	100	100
300	200	8.6	4.3	9.2	6.8	99.3	79.0	100	100	98.4	43.3	100	100
300	250	11.4	4.4	11.9	5.8	99.8	94.3	100	100	99.5	82.6	100	100
300	300	11.3	5.9	12.1	7.7	99.8	99.0	100	100	99.8	96.3	100	100

Notes: The DGP is $X_{it} = \sum_{k=1}^{2} \alpha_{ik} F_{kt} + e_{it}$ where $F_{kt} = \phi_k F_{k,t-1} + v_{kt}$, $\alpha_{ik}, e_{it}, v_{kt} \sim i.i.d\ N(0,1)$, and $[\phi_1, \phi_2] = [0.8, 0.2]$. The shifts in the means of the factor loadings are 0.4 and 0.2 at $\tau = T/2$ (See Section 1.5.3).

The results are reported in Table 1.2 for the Wald and Sup-Wald tests, where two conclusions can be drawn. First, the serial correlations of e_t generally have a larger impact on the sizes than the cross-sectional correlations. Second, the size distortions of the Wald tests disappear as the sample size grows. Overall, the Wald tests have correct sizes for $\bar{r} = 4$ when $T = 200$.[14]

1.5.3 Power properties

We next consider similar DGPs as in Table 1.1 but this time with $r = 2$ and now subject to big breaks which are characterized as deterministic shifts in the means of the factor loadings.[15] The factors are simulated as AR(1) processes with coefficients of 0.8 for the first factor and 0.2 for the second. The shifts in the loadings are 0.2 and 0.4 at time $\tau = T/2$. The other parts of the DGP are the same as in Table 1.1. Table 1.3 reports the empirical powers of the LM/Wald and Sup-LM/Wald tests in percentage terms with 1000 replications. As expected, both tests are powerful to detect the breaks as long as $\bar{r} > r = 2$, while the power is trivial when $\bar{r} = r = 2$.

[14]The only exception is when the errors are all strongly correlated ($\rho = 0.7$ implies that the errors are even more persistent than two of the factors). In practice, this can be tested since e_t can be consistently estimated.

[15]The results with other types of breaks, such as random shifts, are similar and available upon request.

TABLE 1.4: Empirical Power of Nominal 5% Size Tests for Different Choices of $\bar{r}$ when $r = 2$, $k_1 = 1$, and $\hat{F}_{1t}$ is used as the regressand.

N	T	$\bar{r}=2$				$\bar{r}=3$				$\bar{r}=4$			
		LM	Sup LM	Wald	Sup Wald	LM	Sup LM	Wald	Sup Wald	LM	Sup LM	Wald	Sup Wald
50	50	9.7	4.5	64.1	66.5	2.0	3.7	99.0	99.9	1.4	3.5	98.7	100
50	100	13.9	4.3	60.5	64.0	8.7	1.4	98.3	99.5	2.0	0.7	98.2	99.7
50	150	15.3	7.8	56.9	62.1	15.2	4.6	98.4	99.7	8.8	1.2	98.4	99.8
50	200	17.6	11.9	56.5	62.4	19.5	9.4	97.2	100	12.3	3.4	97.6	100
100	50	10.6	4.1	69.6	72.3	3.0	4.6	99.5	100	2.3	3.7	99.5	100
100	100	14.2	5.1	67.1	70.3	9.0	2.1	99.8	100	3.5	0.9	99.8	100
100	150	15.4	10.3	63.1	68.4	13.5	4.0	99.4	100	8.0	1.2	99.2	100
100	200	17.2	11.3	67.5	72.1	16.5	7.7	99.7	100	10.5	2.2	99.7	100
150	50	9.7	5.3	73.9	76.8	2.9	3.9	99.9	100	1.2	2.9	100	100
150	100	12.2	5.3	70.5	74.1	7.8	2.1	99.8	100	1.8	0.7	99.8	100
150	150	12.2	7.3	69.6	74.7	9.3	3.0	99.7	100	5.2	0.9	99.9	100
150	200	12.3	10.7	64.9	70.5	11.7	7.6	99.9	100	6.6	2.2	99.9	100
200	50	9.4	5.8	75.3	77.3	2.9	4.3	100	100	1.9	3.5	100	100
200	100	10.8	5.7	71.6	76.4	8.0	2.1	99.8	100	2.6	0.8	99.9	100
200	150	13.1	10.4	70.9	79.0	9.4	4.7	99.8	100	5.6	0.9	99.9	100
200	200	13.8	10.2	73.1	76.8	13.7	7.4	100	100	8.4	2.0	100	100

Notes: The DGP is $X_{it} = \alpha_{i1}F_{1t} + \alpha_{i2}F_{2t} + e_{it}$ for $t = 1, \ldots, T/2$, and $X_{it} = \alpha_{i1}F_{1t} + \beta_{i2}F_{2t} + e_{it}$ for $t = T/2 + 1, \ldots, T$, where $F_{kt} = \phi_k F_{k,t-1} + v_{kt}$, $\alpha_{i1}, \alpha_{i2}, \beta_{i2}, e_{it}, v_{kt} \sim i.i.d\ N(0,1)$, and $[\phi_1, \phi_2] = [0.8, 0.2]$ (See Section 1.5.3).

Next, we study the powers of our tests when the argument in Section 1.4.3 fails, i.e., the first element of ρ_1 and ρ_2 are both zero. The DGPs are constructed as follows:

$$\begin{aligned} X_{it} &= A_1F_{1t} + A_2F_{2t} + e_{1t} \text{ for } t = 1, \ldots, T/2 \\ X_{it} &= A_1F_{1t} + B_2F_{2t} + e_{1t} \text{ for } t = T/2 + 1, \ldots, T, \end{aligned}$$

where F_{1t} and F_{2t} are two AR(1) process defined as above, $A_1 \sim N(0,1)$, $A_2 \sim 0.9 \cdot N(0,1)$, $B_2 \sim 0.8 \cdot N(0,1)$, and $e_{it} \sim N(0,1)$. Define $F_{2t}^1 = F_{2t}\mathbf{1}(t < \tau)$, $F_{2t}^2 = F_{2t}\mathbf{1}(t \geq \tau)$, and $F_t^* = (F_{1t}' \; F_{2t}^{1'} \; F_{2t}^{2'})'$. Then, if $\bar{r} = 3$, using results of Bai (2003) it is easy to show that:

$$\hat{F}_t = D^* F_t^* + o_p(1) \text{ for } t = 1, \ldots, T,$$

where D^* is a 3×3 diagonal matrix. It then follows from the definition of F_t^* that ρ_1 and ρ_2 should be vectors taking the form $(0, 0, a)$ and $(0, b, 0)$, respectively, with $a, b \neq 0$. Our tests are applied to such DGPs for $\bar{r} = 2, 3, 4$, first using $\hat{F}_{1t}$ as the regressand and then $\hat{F}_{2t}$. The results with 1000 replications are reported in Tables 1.4 and 1.5.

First, it is clear that, even for this special DGP, our Wald and Sup-Wald tests exhibit good power in finite samples when $\hat{F}_{1t}$ is used as the regressand. Second, although our LM and Sup-LM tests lose power when $\hat{F}_{1t}$ is the regressand, this is not the case when $\hat{F}_{2t}$ is used as the regressand. Finally, the Wald and Sup-Wald tests strongly dominates the LM and Sup-LM tests in term of power. Many other simulations have been run in which the power of our Wald and Sup-Wald tests are found to be strong and robust. As will be discussed below, the differences between the power of our LM and Wald tests for this DGP can be

TABLE 1.5: Empirical Power of Nominal 5% Size Tests for Different Choices of $\bar{r}$ when $r = 2$, $k_1 = 1$, and $\hat{F}_{2t}$ is used as the regressand.

N	T	$\bar{r} = 2$				$\bar{r} = 3$				$\bar{r} = 4$			
		LM	Sup LM	Wald	Sup Wald	LM	Sup LM	Wald	Sup Wald	LM	Sup LM	Wald	Sup Wald
50	50	9.7	4.5	20.2	32.1	17.8	2.5	94.4	80.6	3.0	2.7	87.1	80.2
50	100	13.9	4.3	16.6.5	23.3	87.8	2.1	98.9	88.9	78.7.0	0.3	98.6	84.7
50	150	15.3	7.8	18.0	20.9	92.5	78.4	99.7	100	89.9	21.9	99.5	97.0
50	200	17.6	11.9	19.5	22.3	95.1	90.9	99.7	100	93.0	85.9	99.7	100
100	50	10.6	6.1	19.5	33.7	18.8	2.8	96.0	84.8	2.6	3.3	91.6	83.8
100	100	14.2	5.1	19.5	25.1	87.4	1.9	99.6	90.6	79.9	1.4	99.5	87.1
100	150	15.4	10.3	18.3	21.1	93.8	81.6	99.4	100	90.9	25.3	99.2	97.1
100	200	17.2	11.3	19.2	20.6	94.7	90.9	99.6	100	93.3	85.2	99.5	100
150	50	9.7	5.3	17.4	32.2	20.8	2.9	97.1	87.9	3.1	2.8	93.0	86.4
150	100	12.2	5.3	17.1	23.0	88.4	2.1	99.9	93.3	80.3	0.8	99.9	90.6
150	150	12.2	7.3	14.3	23.3	93.3	79.1	99.7	100	91.2	23.0	99.6	98.1
150	200	12.3	10.7	14.7	19.3	94.5	90.7	99.7	100	93.6	85.7	99.7	100
200	50	9.4	5.8	17.7	32.7	20.0	3.9	96.5	87.8	2.4	4.7	92.4	87.0
200	100	10.8	5.7	15.1	24.1	87.4	2.4	100	94.4	78.4	1.1	99.9	90.9
200	150	13.1	10.4	15.3	23.3	93.1	80.7	99.9	100	91.2	26.2	99.8	98.4
200	200	13.8	10.2	15.8	20.3	93.9	89.6	99.7	100	92.3	84.9	99.7	100

Notes: DGP of Table 1.4 (See Section 1.5.3).

explained by the fact that the former uses much less information than the latter. Therefore, we recommend the use of Wald and Sup-Wald on the basis of their good size and power properties.

1.5.4 Comparison with the BE test

As discussed earlier, the BE test relies on the consistent estimation of the original factors. These authors construct N test statistics s_i for each of the hypothesis $\alpha_i = \beta_i$, but not for the joint hypothesis $A = B$. Their method have two limitations. First, as we have shown, the big breaks lead to a new FM representation, as in (1.5), in which the new factor loadings are constant. Thus, the BE test will lose power when the number of factors is chosen to be $r + k_1$, which is quite possible in light of our Proposition 1.2. On the contrary, our tests will not suffer from this problem when $\bar{r} > r$. To compare the performance of our test againts the BE test for the joint hypothesis ($A = B$), we need to construct the following pooled statistic as suggested by Remark A of BE (2011):

$$\frac{\sum_{i=1}^{N} s_i - N\bar{r}}{\sqrt{2N\bar{r}}}$$

where s_i is the individual LM statistics in BE (2011). This test should converge to a standard normal distribution as long as e_{it} and e_{jt} are independent, a restrictive assumption that we also adopt here for comparative purpose. For simplicity, we only report results for the case of known break dates.

We first generate FMs with $r = 2$, and compare the performances of the pooled BE test with our Wald test under the null. The DGPs are similar to those used in the size study. The

TABLE 1.6: Size and Power Comparisons of BE (2011) and Our Wald Tests at Nominal 5% Size for $r = 2$.

N	T	no break, $\bar{r} = 2$		1 break, $\bar{r} = 2$		1 break, $\bar{r} = 3$	
		BE	Wald	BE	Wald	BE	Wald
100	100	6.0	3.9	100	5.6	21.9	96.8
100	150	5.9	5.2	100	7.2	18.2	100
100	200	5.2	4.3	100	6.2	26.0	89.8
100	250	5.3	4.8	100	8.7	17.9	97.7
100	300	5.7	4.3	100	7.4	30.2	83.9
150	100	6.4	4.3	100	5.8	18.3	94.6
150	150	5.9	5.7	100	6.6	16.2	100
150	200	5.6	4.3	100	6.2	12.5	100
150	250	5.5	4.5	100	5.7	14.9	98.3
150	300	4.9	4.0	100	5.6	20.6	89.7
200	100	5.5	4.1	100	4.1	20.0	95.8
200	150	5.4	4.8	100	6.6	15.8	100
200	200	7.0	4.5	100	6.3	14.0	100
200	250	6.5	4.7	100	7.5	12.6	100
200	300	5.0	4.7	100	7.8	12.0	99.7
250	100	6.8	3.9	100	4.2	18.8	97.0
250	150	5.4	5.3	100	5.9	14.9	100
250	200	4.5	4.6	100	6.1	11.3	100
250	250	5.1	4.2	100	6.6	10.9	100
250	300	6.6	4.9	100	8.3	7.9	100
300	100	7.3	4.7	100	5.4	19.7	96.3
300	150	7.0	3.6	100	6.1	14.4	100
300	200	5.9	3.4	100	6.0	13.6	100
300	250	5.9	5.4	100	6.7	12.0	100
300	300	5.7	6.1	100	7.0	10.0	100

Notes: DGP of Table 1.3. The shift in the mean of the factor loadings is either zero (no break) or 0.1 (break) (See Section 1.5.4).

second column of Table 1.5 (no break) reports the 5% empirical sizes. In general, we find that both tests exhibit similar sizes.

Then, we generate a break in the loadings of the first factor while the other elements of the DGPs remain the same as in Table 1.3 where we study the power properties. The break is generated as a shift of size 0.1 in the mean of the loadings. As before, we consider two cases: (i) the number of factors is correctly selected: $\bar{r} = r = 2$; and (ii) the selected number of factors is larger than the true one: $\bar{r} = 3 > r = 2$. The third and fourth columns in Table 1.5 report the empirical rejection rates of both tests. In agreement with our previous discussion, the differences in power are quite striking: when $\bar{r} = 2$, the pooled BE test is much more powerful while the opposite occurs when $\bar{r} = 3$. Notice that, as discussed above, for $\bar{r} = 2$, our test will not be able to detect the break whereas, for $\bar{r} = 3$, the pooled BE test will be powerless. However, according to our Proposition 1.2, the use of BN's (2002) IC will yield the choice of $\bar{r} = 3$ as a much more likely outcome as N and T increase. For example, for this simulation, on average the PC_{p1} of BN (2002) chooses $\widehat{r} = 3$ in 94.6% of the cases.

The second problem of the BE approach is that, even when r is assumed to be known, their method may falsely reject the null hypothesis $\alpha_i = \beta_i$ for some i. To see this, we can use the same argument as in (1.11) to (1.13) to show that, in the presence of big breaks, for some

variables with constant loadings[16], replacing the true factors by the estimated factors will result in a "break" in the factor loadings.

1.5.5 Comparison with the HI test

The HI (2012) test is based on the comparison of the covariance matrices of the estimated factors before and after the break. In view of our results in (1.9) and (1.10), $\tau^{-1}\sum_{t=1}^{\tau}\hat{F}_t\hat{F}_t' = D_1\hat{\Sigma}_F^1 D_1' + o_p(1)$, and $(T-\tau)^{-1}\sum_{t=\tau+1}^{T}\hat{F}_t\hat{F}_t' = D_2^{\dagger}\hat{\Sigma}_F^2 D_2^{\dagger'} + o_p(1)$, where $\hat{\Sigma}_F^1 = \tau^{-1}\sum_{t=1}^{\tau}F_tF_t'$ and $\hat{\Sigma}_F^2 = (T-\tau)^{-1}\sum_{t=\tau+1}^{T}F_tF_t'$. Therefore, the HI test is able to detect breaks if $\hat{\Sigma}_F^1, \hat{\Sigma}_F^2 \to \Sigma_F$ and D_1 and $D_2^{\dagger}$ converge to different limits as N and T go to infinity. Specifically, their test is defined as:

$$T\left(\mathcal{C}(\pi)'\hat{V}^{-1}\mathcal{C}(\pi)\right)$$

where

$$\mathcal{C}(\pi) = \text{vech}\left[\frac{1}{\tau}\sum_{1}^{\tau}\hat{F}_t\hat{F}_t' - \frac{1}{T-\tau}\sum_{\tau+1}^{T}\hat{F}_t\hat{F}_t'\right],$$

and $\hat{V}$ is a HAC estimator of the covariance matrix of $\mathcal{C}(\pi)$ which is either constructed using the whole sample (LM version of the test) or using subsamples before and after the break (Wald version).

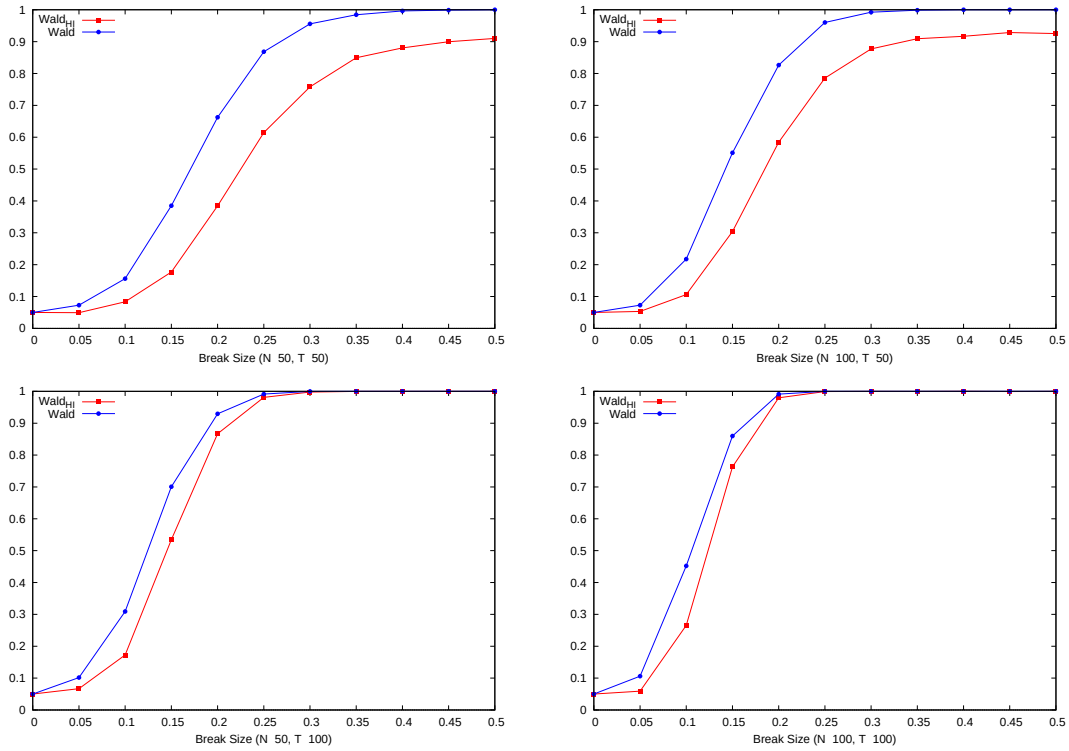
Basically, the HI test exploits the same insight as our tests in converting an infinite-dimensional problem to a finite one, except that it relies on a different use of the estimated factors. Compared to ours, the HI test uses more information since our LM test only uses the first row (except the first element) of $\frac{1}{\tau}\sum_{t=1}^{\tau}\hat{F}_t\hat{F}_t'$, while our Wald test uses all the elements of the matrix except the first one ($\frac{1}{\tau}\sum_{1}^{\tau}\hat{F}_{1t}^2$).

In principle, it may seem that the use of less information is the price one has to pay to render our testing procedure much simpler than theirs. After all, both steps in our approach can be easily implemented in any conventional statistical software used by practitioners, while HI's test is computationally more burdensome. Yet, our Wald test exhibits very similar power to theirs in all the simulations we have run. HI (2012) reports some simulation results for the power comparisons with our Wald tests under very general DGPs. Therefore, to avoid repetitions, we focus only on small samples ($N, T \leq 100$), and compare the (size-adjusted) power curves of HI's and our Wald tests (using the Bartlett kernel) for the following DGP: $X_{it} = A_{1i}F_t + e_{it}$ for $t = 1, \ldots, T/2$, and $X_{it} = (A_{1i} + b)F_t + e_{it}$ for $t = T/2+1, \ldots, T$, where $F_t = \rho F_{t-1} + u_t$, $A_{1i}, u_t, e_{it} \sim i.i.d\ N(0,1)$, $\rho = 0.8$, and b is the break size which ranges from 0 to 0.5. As can be observed in Table 1.7, our Wald test has better power properties than HI's test in all these cases. However, not surprisingly, as N and T get large, both tests perform very similarly in term of power.[17]

[16] This is possible since our definition for big breaks allows a fixed proportion of variables to have no breaks.

[17] This can be seen from Tables 5A and 5B of HI (2012). We also have similar unreported simulation results that are available upon request.

TABLE 1.7: Power Comparison of HI (2012) and Our Wald Tests for $N, T \leq 100$ when $r = 1$.



Notes: The DGP is $X_{it} = A_{1i}F_t + e_{it}$ for $t = 1, \ldots, T/2$, and $X_{it} = (A_{1i} + b)F_t + e_{it}$ for $t = T/2 + 1, \ldots, T$, where $F_t = \phi F_{t-1} + u_t$, $A_{1i}, u_t, e_{it} \sim i.i.d\ N(0, 1)$, and $\phi = 0.8$. The reported curves are the size adjusted power curves of our Wald test (blue) and HI's Wald test (red) when the break size b increases from 0 to 0.5 (See Section 1.5.5).

1.6 An Empirical Application

To provide an empirical application of our tests, we use Stock and Watson's (2009) data set consisting of 144 quarterly time series of nominal and real variables for the US ranging from 1959:Q1 to 2006:Q4. Since not all the variables are available for the whole period, we end up using their suggested balanced panel of standardized variables with $T = 190$, $N = 109$. This more or less corresponds to the case where $T = 200$, $N = 100$ in Tables 1.1 and 1.2, where no severe size distortions are found. We refer to Stock and Watson (2009) for the details of the the data description and the standardization procedure.

Using various BN's (2002) IC (IC_{p1}, IC_{p2}, PC_{p1} and PC_{p2}) the estimated number of factors ranges from 2 to 6, therefore we implement our test for $\bar{r} = 2$ to 6. The Sup-Wald test is applied since no priori break date is assumed to be known. In order to have enough observations in both subamples, we use the trimming $\Pi = [0.3, 0.7]$. It corresponds to the time period ranging from 1973:Q3 to 1992:Q3 which includes several relevant events like, e.g., the oil price shocks (1973, 1979) and the beginning of great moderation period in the early

1980s. The graphs displayed in Figure 1.2 are the series of Wald tests for different values of $\bar{r}$, with the horizontal lines representing the 5% asymptotic critical values of the Sup-Wald tests.

We find that the Sup-Wald test rejects the null when $\bar{r} = 5, 6$. The estimated break date is around 1979-1980 (second oil price shock), rather than 1984, which is the only candidate considered by Stock and Watson (2009) as a potential break date in their empirical application with the same data set. One possible explanation for this break date could be the Iranian revolution at the beginning of 1979 and its subsequent impact on monetary policy in the US (see Fernandez-Villaverde et al., 2010).

1.7 Conclusions

In this paper, we propose a simple two-step procedure to test for big structural breaks in the factor loadings of large FM that circumvents some limitations affecting other available tests. In particular, after choosing the number of factors in the whole sample according to BN's (2002) IC and estimating them by PCA, our test relies on a regression of one of the estimated factors on the remaining ones, allowing for a break in the parameters at known or unknown date. LM and Wald tests for the null of parameter stability are applied. We show that our test may have better power than the test of BE (2011) under the alternative of big breaks and, and that it is simpler than the test of HI (2012). Despite using less information than the latter, we show that it is used in a more efficient way, and that our Wald test has better power properties than the HI test when dealing with small samples.

Our testing approach can be easily implemented in any statistical package and it is is useful to avoid serious forecasting/estimation problems in standard econometric practices with factors. This could be the case of FAR and FAVAR models, when the factor loadings are subject to big breaks and the number of factors is a priori determined (as is conventional in several macroeconomic and financial applications).

In the second step of our testing approach, a Sup-type test is used to detect a break of the parameters in that regression when the break date is assumed to be unknown. As the simulations show, this test performs very well especially when $T \geq 100$. For smaller samples, as it happens with many other Sup-type tests, bootstrap can improve the finite-sample performance of the test compared to the tabulated asymptotic critical values of Andrews (1993), as suggested by Diebold and Chen (1996). It is high in our research agenda to explore this possibility.

Moreover, as discussed earlier in Remarks 8 and 9, many other existing methods for testing structural changes in linear regressions can also be applied in our second-stage procedure. Further, our testing approach can allow for multiple big breaks through sequential estimation,

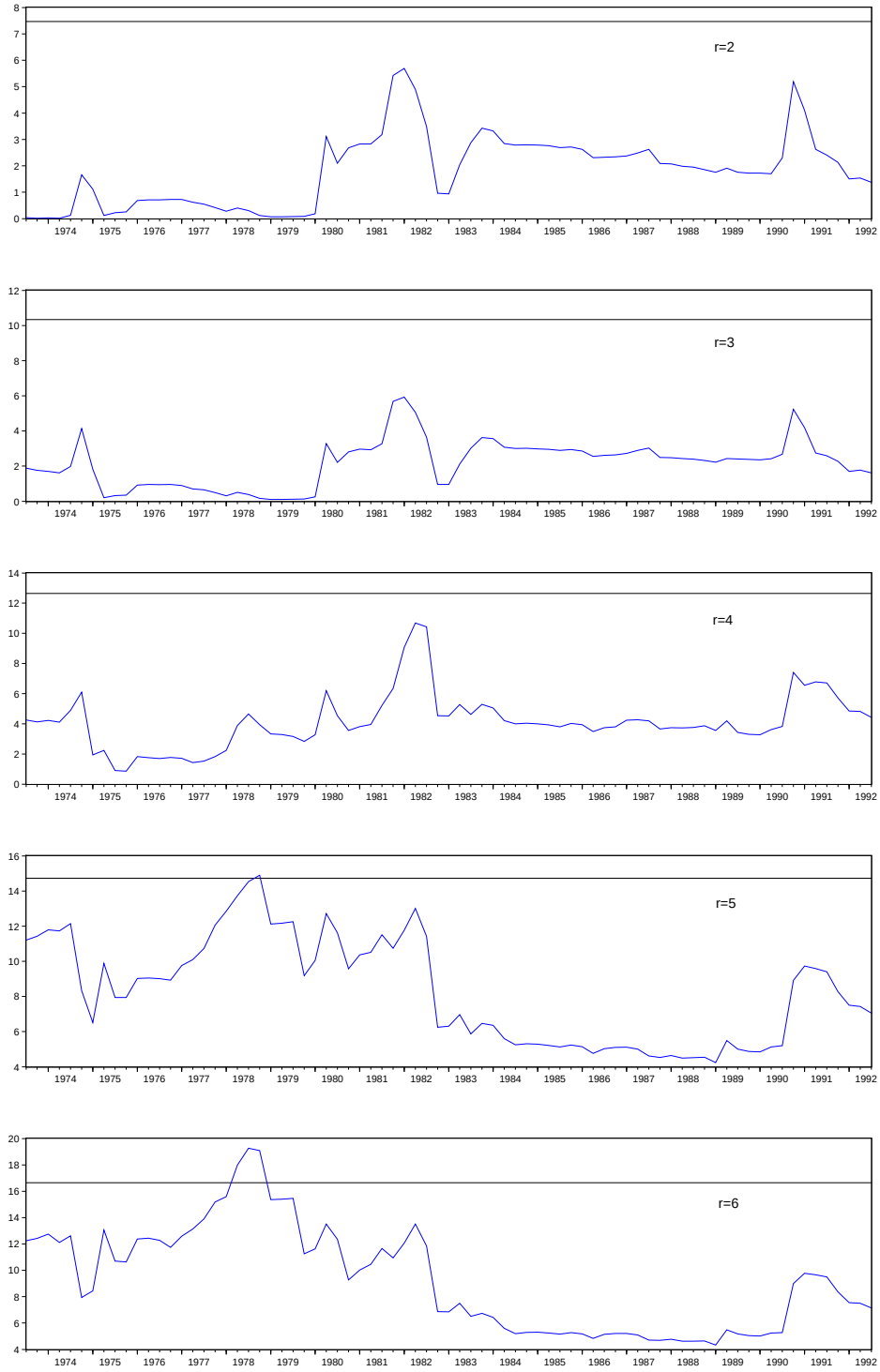


FIGURE 1.2: US data set of Stock and Watson (2009), from 1959:Q1 to 2006:Q4. The trimming $\Pi = [0.3, 0.7]$ is used for the Wald tests with $\bar{r} = 2$ to 6 (from top to bottom), and the horizontal lines are the corresponding 5% asymptotic critical values for the Sup-Wald Test.

like in Bai and Perron (1998, 2003), for locating the candidate break dates. Exploring further this issue remains also high in our research agenda.

Finally, though a simple testing procedure has been outlined in Section 1.4.4 to identify whether breaks stem from loadings or from the volatility of the factors, we plan to derive other alternative tests based on the rank of the covariance matrix of the estimated factors in different subsamples which can also be extended to test for other sources of parameter instability.

Chapter 2

Identifying Observed Factors in High Dimensional Factor Models

2.1 Introduction

Factor models (FM henceforth) are becoming an increasingly important tool for both theoretical and empirical research. For example, in macroeconomics, the solutions of Dynamic Stochastic General Equilibrium (DSGE) models can be written in the form of FM when these models allow for measurement errors (Altug 1989 and Sargent 1989), so that the structure of FM can help solve these models even when a large number of variables are considered (Boivin and Giannoni 2006, Kryshko 2011). In structural analysis, the factors estimated from large panel datasets can be combined with Structural Vector Autoregressions (SVAR) to identify the effects of fundamental shocks (Bernanke et al 2005), and solve the problem of *non-fundamentalness* (Forni et al 2009). Moreover, the estimated factors can significantly improve the forecasts of macro variables (Stock and Watson 2002a). In microeconomics, the demand systems are shown to have a factor structure (Lewbel 1991), and in some recent studies, FM are used to characterize the unobservable cross-sectional dependencies in panel data models (Pesaran 2006 and Bai 2009). Finally, in finance, the key assumption underlying the Arbitrage Pricing Theory (APT) is the multi-factors structure for the asset returns.

The popularity of FM is mainly due to their capability of summarizing co-movements of a large number of variables (N) by a much smaller number of common factors ($r << N$). Moreover, the rapidly increasing dimensions of available data sets allow us to depart from the restrictive assumptions of the classical factor analysis, and estimate the factor models consistently using the method of Principal Components (PC hereafter) (Bai and Ng 2002, Bai 2003, Stock and Watson 2002a).

Yet, it is well recognized that FM suffer from identification problems. Consider a factor model: $X_t = \Lambda f_t + e_t$, where X_t is the vector of observed variables, Λ is the matrix of factor

loadings, f_t is the unobservable factors, and e_t is the vector of idiosyncratic errors. Since only X_t can be observed, the above model is observably equivalent to: $X_t = (\Lambda H^{-1})(Hf_t) + e_t$, where H is any $r \times r$ nonsingular matrix. Therefore, unless one imposes $r \times r$ prior restrictions, the factors can only be identified up to a rotation, and thus the estimated factors usually lack a direct interpretation.[1]

In some situations, the object of interest is the conditional mean of some observed variables, so that the interpretation of the factors is not important. For example, in panel data models, one only needs to consistently estimate the common parts ($\mathbf{\Lambda} f_t$) of the unobservable effects, and thus the indeterminacy of the factors rotation does not matter for the results.

However, there are other instances where the direct object of interest are the factors themselves and thus a clear interpretation of them can have important implications for structural analysis. In financial economics, a large body of empirical research is concerned with identifying the factors that determine asset returns. Chen et al (1986) and Shanken and Weinstein (2006) are examples of such work that try to interpret the underlying forces in the stock market in terms of some observed macro variables. Instead of using macro variables, Fama and French (1993) identify three observed factors related to the market returns and firm characteristics, which can explain a large proportion of return volatilities. In the solutions of DSGE models, the state variables and exogenous shocks (e.g., preference shocks or technology shocks) play the role of common factors, so that the interpretation of the factors is equivalent to identifying the sources of business cycles. In factor-based forecasts, not all the estimated factors necessarily have prediction power for the target variables, and hence the forecasting can be further improved if some interpretational contents are attached to the factors. For example, the predictions of inflation rates could be more accurate if the factors associated with monetary policy shocks are given more weight than other factors identified as productivity changes.

The goal of this paper is to identify the factors by relating them to some observed variables. The point of departure is the assumption that the common factors can be well approximated by (or linear functions of) some observed variables. Under this assumption, we will denote these observed variables as *observed factors.* We focus on the approximate factor models (Chamberlain and Rothschild 1983, Bai and Ng 2002) which allow for quite general assumptions about the data generating processes (DGP henceforth). More importantly, the space of the factors can be consistently estimated using the method of PC under the assumption of large N (the number of variables).

To the best of our knowledge, Bai and Ng (2006) is the only work that has addressed this issue.[2] They consider the null hypothesis: $g_t = Lf_t$ for a $m \times r$ matrix L and a list of $m(> r)$

[1]The conventionally adopted identification assumptions for the estimation of factors using PC are that: (i) the factors are orthogonal and (ii) the covariance matrix of the factor loadings is diagonal.

[2]Bai and Ng (2011) study the identification of factors from a statistical point of view, i.e., by imposing restrictive assumptions on the data generating processes of the factors and factor loadings.

observed variables g_t, suggested by some economic reasoning. They develop test statistics for each of the observed variables g_{kt} as well as for the whole set of variables g_t, based on the regressions of g_t on the estimated factors.

In practice, however, the list of observed factors is not always available, or those suggested by economic theory may not span the same space of the underlying factors. In view of these caveats, we propose here to first select a list of observed factors from a much larger set of variables, and then test the null hypothesis that the underlying factors are exact linear combinations of observed variables selected in the first step.

In the first part, the estimated factors are regressed on different subsets of observed variables, and we label as the *estimated observed factors* the subset of variables that minimizes the Residual Sum of Squares (RSS) in these regressions. We differentiate two cases of observed factors: the *directly observed factors* (DOFs henceforth) and the *indirectly observed factors* (IOFs). In the first case, the latent factors in the FMs are directly approximated by the observed factors, i.e., there is a one-to-one correspondence between the r factors and r observed variables. In the second case, by contrast, the r factors are linear functions of m observed variables with $m \geq r$. Notice that this second setup includes the first one as a special case, but we will show that, for DOFs, our estimation method is much simpler and allows for larger measurement errors (i.e., the difference between the latent factors and the observed factors).

The above methods, known as *subset search regressions* in model selections, are shown to be consistent under some general assumptions. However, their computation costs can be huge or even unaffordable when the number of candidate variables is large. We thus consider a much more efficient procedure called the Lasso (Least Absolute Shrinkage and Selection Operator). In such a procedure, the estimated factors is again regressed on a large number of observed variables, but a special penalty function is added to the object function so that only a few variables will be selected. We show that a modified version of the Lasso is able to select the observed factors consistently, but under more stringent assumptions than those of the subset search methods.

In the testing part, we consider the null hypothesis: $f_t = BX_{1:m,t}$ for a $r \times m$ matrix B and a list of m observed variables $X_{1:m,t}$. This hypothesis covers both cases (DOF and IOF) discussed above, and is shown to be more general than the hypothesis considered by Bai and Ng (2006). We derive two types of test statistics based on the residuals in the regressions of estimated factors on $X_{1:m,t}$.

Our paper contributes to the literature of FM in 2 dimensions. First, we provide some practical methods to identify the factors that can be quite general functions of some observed variables, e.g., the innovations in an AR model for consumption growth. These methods efficiently explore the information of a large data set and will consistently identify the observed factors. Our results also formalize the R^2 based method widely used in practice by providing

rigors theoretical proof. Second, we provide several interesting applications in finance and macroeconomics where our methods are shown to be able to identify some observed factors. We confirm that the Fama-French 3 factors are good proxies for the underlying risk factors, not only for portfolios returns but also for stock returns that have much larger idiosyncratic errors.

The rest of the paper is organized as follows: Section 2.2 defines the basic notations and discusses the assumptions that define the approximate factor models. In Section 2.3, we show how to identify the observed factors using regression-based methods discussed above. In Section 2.4, we define the null hypothesis for the observed factors and propose several test statistics whose asymptotic distributions are also derived. Section 2.5 studies the finite sample properties of the estimation methods and the test statistics. In Section 2.6 we apply our method to identify the factors for the returns of stocks and portfolios, and for a large panel of macroeconomic series. Finally, Section 2.7 concludes. All the proofs are collected in the Appendices.

2.2 Models, Notations and Assumptions

Throughout this paper, we use the following standard notation. We define the matrix norm: $\|A\| = \sqrt{\text{Tr}(A'A)}$, and use $A_{1:m}$ to denote the 1st to mth rows of a matrix (or a vector) A. Further, $A > 0$ (≥ 0) means that the matrix A is positive (semi) definite.

The following approximate factor models are considered:

$$X_t = \Lambda f_t + e_t, \tag{2.1}$$

where $X_t = (x_{1t}, \ldots, x_{Nt})'$ is a $N \times 1$ vector of observed variables, $\Lambda = (\lambda_1, \ldots, \lambda_N)'$ is a $N \times r$ vector of factor loadings, $f_t = (f_{1t}, \ldots, f_{rt})'$ is a $r \times 1$ vector of common factors, and $e_t = (e_{1t}, \ldots, e_{Nt})'$ is a $N \times 1$ vector of idiosyncratic errors. Unlike the classical factor analysis, we allow the number of variables N to go to infinity and the errors $\{e_{it}\}$ to be both temporarily and cross-sectionally correlated.

We assume that m among the observed variables X_t are *observed factors*, in a sense to be defined in the following sections, where m is a fixed number that does not increase as N goes to infinity. Without loss of generality, we assume these m *observed factors* are ordered as the first m variables of X_t. The main issue is how to find these m observed variables in the available data set of size N . Given that the m observed factors are always placed in the first m rows, this issue becomes equivalent to finding out the first m variables out of N randomly ordered variables X_t.

We consider two cases: DOFs and IOFs. In either case, the m observed factors have the following form:

$$X_{1:m,t} = \Lambda_{1:m} f_t + e_{1:m,t}. \tag{2.2}$$

To single out the observed factors, we have to impose some restrictions on $\Lambda_{1:m}$ and $e_{1:m,t}$, which will be discussed in the next section. Roughly speaking, for the DOFs, $\Lambda_{1:m}$ should be a full rank matrix and $e_{1:m,t}$ should go to zero as N and T go to infinity; for IOFs, a necessary condition is that the covariance matrix of $e_{1:m,t}$ has reduced rank. These restrictions allow us to write:

$$f_t = BX_{1:m,t} + o_p(1),$$

therefore the true factors can be closely approximated by linear combinations of the observed factors.

Next we impose some assumptions for Λ, f_t and e_t, which are necessary for the consistency of the estimated factors using PC. Further, it should be noted that the assumptions to be imposed in the next sections, when defining the observed factors, do not contradict with the following ones.

Let M denote a finite constant, we assume that:

Assumption 11. *(i)* $E||f_t||^4 < M$ *for* $t = 1, \ldots, T$, *and* $T^{-1}\sum_1^T f_t f_t' \xrightarrow{p} \Sigma_F > 0$ *as* $N, T \to \infty$; *(ii)* $||\lambda_i|| < M$ *for* $i = 1, \ldots, N$, *and* $\left\|N^{-1}\sum_1^N \lambda_i \lambda_i' - \Sigma_\Lambda\right\| = O_p(1/\sqrt{N})$ *for some* $\Sigma_\Lambda > 0$; *(iii) The* r *eigenvalues of* $\Sigma_\Lambda \Sigma_F$ *are different.*

Assumption 12. *(i)* $E(e_{it}) = 0$, $E(e_{it})^8 \leq M$;
(ii) For $i, j = 1, \ldots, N$ *and* $s, t = 1, \ldots, T$, $E(e_{it}e_{js}) = \tau_{ij,ts}$, $|\tau_{ij,ts}| \leq \tau_{ij}$ *for all* (t, s), *and* $|\tau_{ij,ts}| \leq \gamma_{ts}$ *for all* (i, j). $N^{-1}\sum_{i,j}\tau_{ij} \leq M$, $T^{-1}\sum_{t,s}\gamma_{ts} \leq M$, $(NT)^{-1}\sum_{i,j,t,s}|\tau_{ij,st}| \leq M$, *and* $\sum_s \gamma_{st}^2 \leq M$;
(iii) For any (t, s), $E|N^{-1/2}\sum_{i=1}^N [e_{is}e_{it} - E(e_{is}e_{it})]|^4 \leq M$.

Assumption 13. $\{f_t\}$ *and* $\{e_{it}\}$ *are two independent groups.*

These Assumptions are quite general in the sense that they allow heteroskedasticitiy, temporal and cross-sectional correlations in the factors and idiosyncratic terms. For more discussion on these Assumptions, see Bai (2003). Under Assumptions 11 to 13, the Information Criteria (IC) proposed by Bai and Ng (2002) can consistently estimate the number of factors, so that we can proceed as if this number was known. The effect of misspecification of the factor numbers will be discussed in the next section.

Define the $T \times r$ matrix $\tilde{F} = (\tilde{f}_1, \ldots, \tilde{f}_T)'$ as $\sqrt{T}$ times the eigenvectors corresponding to the r largest eigenvalues of the $T \times T$ matrix XX', where the $T \times N$ matrix $X = (X_1, \ldots, X_T)'$. Then, denoting $\min[\sqrt{N}, \sqrt{T}]$ as $\delta_{N,T}$, the following result holds:

Lemma 2.1. *(Bai and Ng 2002) Under Assumptions 11 to 13,* $\delta_{N,T}||\tilde{f}_t - Hf_t|| = O_p(1)$ *for* $t = 1, \ldots, T$, *where* $H' = \left(\Lambda'\Lambda/N\right)\left(F'\tilde{F}/T\right)V_{NT}^{-1}$, *and* V_{NT} *is a diagonal matrix containing the* r *largest eigenvalues of* $(NT)^{-1}XX'$.

Lemma 2.1 is the key result underlying our identification method for observed factors. It implies that the estimated factors are consistent for the space spanned by the true factors and hence for the observed factors. This relationship between the estimated factors and observed factors can be explored to identify the latter. For the identification of IOFs, the convergence rate $\delta_{N,T}$ is important to design an appropriate penalty function to consistently estimate m.

However, Assumption 11 excludes the *weak factors* as in Onatski (2009a), in which the PC estimator of the factors may not be consistent if their explanatory power is small relative to the idiosyncratic terms. For example, in the study of stock returns, it is shown that the volatilities of excess returns are mainly due to idiosyncratic errors (Goyal and Santa-Clara 2006). The identification of observed factors in such weak factor models will also be discussed.

2.3 Identifying Observed Factors Using Regressions

In this section, we show how to identify the observed factors using simple linear regressions of $\tilde{f}_t$ on $X_{1:m,t}$. Given that $\tilde{f}_t$ are consistent for the space of f_t, and that f_t can be approximated by linear combinations of $X_{1:m,t}$, linear projection of $\tilde{f}_t$ on $X_{1:m,t}$ should result in small residuals and high R^2.

However, in many empirical studies using factor models, instead of regressing $\tilde{f}_t$ on $X_{1:m,t}$, each observed variables in X_t are regressed on a single or a set of estimated factors, and then explanations are given to these estimated factors according to R^2 in such regressions. For example, in Ludvigson and Ng (2009), the first estimated factor from a large panel of macro variables is labeled as a *real factor*, because the marginal regressions of the real variables (productions, employment...) on this factor produce relatively large R^2 compared to other nominal variables.

In Section 2.3.1, we formalize these R^2-based methods for the case of DOFs, and show that both approaches (regressing $\tilde{f}_t$ on subsets of X_t and regressing X_t on subsets of $\tilde{f}_t$) can identify the observed factors. In Section 2.3.2, we generalize these methods for IOFs, and show that only regressions of $\tilde{f}_t$ on subsets of X_t can identify all the observed factors. A class of penalized regressions called the Lasso have good model selection properties, and we show they can be used to select observed factors in Section 2.3.3. Section 2.3.4 discusses models with weak factors.

2.3.1 Directly Observed Factos

In this section, we deal with the identification of the DOFs. To give the precise definition of DOFs, the following assumptions are imposed:

Assumption 14. *(i)* $m = r$, $\Lambda_{1:r}$ *has full rank, and* $e_{it} = \kappa_{N,T}\varepsilon_{it}$ *for* $i = 1, \ldots, r$, *where* $\kappa_{N,T} \to 0$ *as* $N, T \to \infty$, *and* $T^{-1}\sum_{t=1}^{T}(\varepsilon_{it}\varepsilon_{jt}) = O_p(1)$ *for* $i, j = 1, \ldots, r$.
(ii) Let $e_{n_1:n_r,t} = (e_{n_1,t}, \ldots, e_{n_r,t})'$ *for* $r + 1 \leq n_1 < n_2 \ldots < n_r \leq N$,
then $T^{-1}\sum_{t=1}^{T} e_{n_1:n_r,t}e'_{n_1:n_r,t} \xrightarrow{p} \Sigma^{e}_{n_1:n_r} > 0$.

Assumption 14(i) states that the first r variables span the space of the common factors f_t asymptotically: $X_{1:r,t} \to \Lambda_{1:r}f_t$ as $N, T \to \infty$. When $\Lambda_{1:r} = I_r$, it simply means the common factors are directly measured by the first r observed variables with neglectable measurement errors. Notice that the nonsingular matrix $\Lambda_{1:r}$ is just a normalization, and hence we can define the new factors as $X_{1:r,t} = \Lambda_{1:r}f_t$, because for the remaining variables we have:

$$\begin{aligned} X_{r+1:N,t} &= \Lambda_{r+1:N}f_t + e_{r+1:N,t} \\ &= \left(\Lambda_{r+1:N}\Lambda_{1:r}^{-1}\right)\left(\Lambda_{1:r}f_t\right) + e_{r+1:N,t} \\ &= \Lambda^{*}_{r+1:N}X_{1:r,t} + e_{r+1:N,t} \end{aligned}$$

Therefore, we label the first r observed variables as *Directly Observed Factors*. Notice Bai and Ng (2006) identify the observed factors by constructing some test statistics under the assumption of an exact relationship between the observed variables and the factors, i.e., $e_{1:r,t} = 0$ for $t = 1, \ldots, T$. By contrast, we allow for small measurement errors — the assumption that $e_{1:r,t} = o_p(1)$ is only a asymptotic approximation for small errors in finite samples.

Assumption 14(ii) rules out (asymptotic) multi-collinearity between any set of r observed variables, such that $T^{-1}\sum_{t=1}^{T} X_{n_1:n_r,t}X'_{n_1:n_r,t} \xrightarrow{p} \Sigma^{x}_{n_1:n_r} > 0$ for $1 \leq n_1 < \ldots < n_r \leq N$, and therefore the DOFs are uniquely defined.

From Lemma 2.1 and Assumption 14 we can derive an approximate linear relationship between the estimated factors and the DOFs:

$$\tilde{f}_t = Hf_t + o_p(1) = H\Lambda_{1:r}^{-1}X_{1:rt} + o_p(1) = AX_{1:r,t} + o_p(1), \tag{2.3}$$

where $A = H\Lambda_{1:r}^{-1}$. As will be defined shortly, our method of identification is based on the regressions of the r estimated factors on r observed variables (in contrast to Bai and Ng 2006 where the observed variables are regressed on the estimated factors). The intuition for our approach is that, if $\tilde{f}_t$ are regressed on the right set of observed variables: $X_{1:m,t}$, the OLS estimator $\hat{A}$ will converge to A and the residuals will be $o_p(1)$, so that RSS$/T$ will converge to 0. If the regressors are chosen as a set of r observed variables different from $X_{1:r,t}$, we show that RSS$/T$ will instead converge to some positive numbers. As a result, we can identify the DOFs by comparing the RSS in the regression of $\tilde{f}_t$ on different sets of observed variables.

Let $n_1 : n_r = [n_1, \dots, n_r]$ denote a set of r indices such that $1 \leq n_1 < n_2 < \dots n_r \leq N$, and let $X_{n_1:n_r,t} = \Lambda_{n_1:n_r} f_t + e_{n_1:n_r,t}$ be the corresponding observed variables. By defining:

$$S(n_1 : n_r, A) = \frac{1}{T} \sum_{t=1}^{T} \left\| \tilde{f}_t - A X_{n_1:n_r,t} \right\|^2, \tag{2.4}$$

and

$$[\hat{n}_1, \hat{n}_2, \dots, \hat{n}_r] = \arg\min_{n_1:n_r} \Big(\min_{A} S(n_1 : n_r, A) \Big), \tag{2.5}$$

then $X_{\hat{n}_1:\hat{n}_r,t}$ is the vector of DOFs identified by our method.

Notice that

$$\frac{1}{T} \sum_{t=1}^{T} \left\| \tilde{f}_t - A X_{n_1:n_r,t} \right\|^2 = \frac{1}{T} \sum_{k=1}^{r} \sum_{t=1}^{T} \left(\tilde{f}_{kt} - A_k X_{n_1:n_r,t} \right)^2,$$

and therefore

$$\min_{A} S(n_1 : n_r, A) = S(n_1 : n_r, \hat{A}),$$

where $\hat{A} = [\hat{A}_1', \hat{A}_2', \dots, \hat{A}_r']'$, and $\hat{A}_k$ is the OLS estimator of A_k. This procedure can be simply implemented as follows: we first choose r observed variables, then calculate RSS_k in the OLS regression of $\tilde{f}_{kt}$ on these chosen variables, and get $\text{RSS} = \sum_{k=1}^{r} \text{RSS}_k$, where the set of variables that yield the smallest RSS are the identified DOFs.

The following theorem states that, using our method, the probability of correctly identifying the DOFs goes to 1 as N and T go to infinity.

Theorem 2.2. *Under Assumptions 11 to 14,* $\mathbb{P}\Big([\hat{n}_1, \hat{n}_2, \dots, \hat{n}_r] = [1, 2, \dots, r] \Big) \to 1$ *as* $N, T \to \infty$.

This result holds as long as $\kappa_{N,T} = o(1)$. However, with finite samples, the DOFs may not be easily distinguishable from the remaining variables, due to either large measurement errors ($\kappa_{N,T}$) or large estimation errors of the PC. The finite sample properties of our identification procedure will be studied in Section 2.5 using simulations.

It is also very easy to see that the DOFs can be identified by regressing variables in X_t on $\tilde{f}_t$, because the RSS/T in the regression of x_{it} on $\tilde{f}_t$ will converge to 0 only when $1 \leq i \leq r$, i.e., the variable x_{it} is one of the DOFs. Thus the r variables that give the smallest RSS will be identified as the DOFs. This approach is commonly used is practice, and its computation costs are much lower than (2.5). However, in the case of IOFs, an extension of (3.3) will correctly identify the observed factors, while regressing X_t on $\tilde{f}_t$ may give misleading results.

2.3.2 Indirectly Observed Factors

2.3.2.1 Definitions and comparison with Bai and Ng (2006)

In the previous section, we have studied the simple case where the common factors are directly observed, i.e., $f_t = X_{1:r,t}$ for $t = 1, \ldots, T$. However, in practice it is quite likely that the common factors are well approximated by the linear combinations of some observed variables, i.e., $f_t = BX_{1:m,t}$ for a $r \times m$ matrix B with full row rank. For example, one of the macro variables considered by Chen et al (1986) is the spread of interest rates. In Cochrane and Piazzesi (2005), a factor that can predict excess bond returns with R^2 up to 0.44 is a linear combination of 5 forward rates. When $m = r$, we have shown in the previous section that this case is equivalent to DOFs. Further, when $m < r$, the space spanned by the factors has rank m, instead of r, and so we should get m factors using Bai and Ng's IC. Hence, without loss of generality, we focus on the case $m > r$ throughout this section.

We impose the following assumption to define the IOFs:

Assumption 15. *(i) $f_t = BX_{1:m,t}$ for $t = 1, \ldots, T$, the $r \times m$ matrix $B = (B_1, \ldots, B_m)$ has full row rank, and $\|B_k\|^2 \neq 0$ for $k = 1, \ldots, m$;*
(ii) For any constant number k, and any set of indices $1 \leq n_1 < \ldots < n_k \leq N$, $T^{-1} \sum_{t=1}^{T} X_{n_1:n_k} X'_{n_1:n_k} \xrightarrow{p} \Sigma^x_{n_1:n_k} > 0$.
(iii) For any set of k indices: $m + 1 \leq n_1 < \ldots < n_k \leq N$, $T^{-1} \sum_{t=1}^{T} e_{n_1:n_k,t} e'_{n_1:n_k,t} \xrightarrow{p} \Sigma^e_{n_1:n_k} > 0$.

The above assumptions guarantee that the set of IOFs $X_{1:m,t}$ are uniquely determined, i.e., there exits no other variables whose linear combinations can span the true factor space. For example, if $X_{1:m,t}$ are IOFs, $X_{1:m+1,t}$ will also be IOFs, since $f_t = (B, \mathbf{0})X_{1:m+1,t}$, but these undesirable cases are excluded by Assumption 15(i).

Note that the assumption $f_t = BX_{1:m,t}$ is essential. To see this, recall that the hypothesis of interest in Bai an Ng (2006) is that $g_t = Lf_t$ for a $m \times r$ matrix L, so that their tests are based on the regressions of the observed variables g_t on the estimated factors $\tilde{f}_t$. On the contrary, as mentioned above, we regress the estimated factors on the observed variables. The difference is trivial for the case of DOFs but not for the case of IOFs. We use a simple example to illustrate this point. Consider a factor model with only one factor: $f_t = x_{1t} - x_{2t}$ for $t = 1, \ldots, T$, where x_{1t} and x_{2t} are two observed variables. The null hypothesis considered by Bai and Ng (2006) is:

$$\begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix} = \begin{pmatrix} c \\ c - 1 \end{pmatrix} f_t, \tag{2.6}$$

where c is any real number. Suppose now there is an estimator $\tilde{f}_t$ such that $f_t = \tilde{f}_t + o_p(1)$, one can write

$$x_{1t} = c\tilde{f}_t + o_p(1), \tag{2.7}$$

and the residuals in the regression of x_{1t} on $\tilde{f}_t$ will be $o_p(1)$ (note that the result is similar for x_{2t}). Their test statistics are based on exploring the exact order of the $o_p(1)$ term, namely $O_p(1/\sqrt{N})$ when $\sqrt{N}/T \to 0$. Now suppose there is another observed variable: $x_{3t} = f_t + e_{3t}$ with $\text{Var}(e_{3t}) = \sigma^2 > 0$. Then, the residuals in the regression of x_{3t} on $\tilde{f}_t$ will be larger than $o_p(1)$ because we can write $x_{3t} = \tilde{f}_t + e_{3t} + o_p(1)$, and their tests have power to reject x_{3t} as a member of g_t.

However, equation (2.6) implies that both x_{1t} and x_{2t} can serve as the true factor and therefore are linearly dependent. Nevertheless, multicollinearity is not very common in practice, and can be avoided by pre checking the data. Moreover, either x_{1t} or x_{2t} can be viewed as DOF, but the assumption considered by Bai and Ng (2006) doesn't include the case of IOFs.

In our definition, only $f_t = x_{1t} - x_{2t}$ is required, whereas x_{1t} and x_{2t} are allowed to have the following representation:

$$\begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix} = \begin{pmatrix} c \\ c-1 \end{pmatrix} f_t + \begin{pmatrix} 1 \\ 1 \end{pmatrix} \varepsilon_t, \tag{2.8}$$

for any real number c and random process ε_t. Note that (2.6) is a special case of (2.8) with $\varepsilon_t = 0$. But with x_{1t} and x_{2t} being defined as in (2.8), the tests of Bai and Ng (2006) will reject the null for both x_{1t} and x_{2t}, despite being true that $f_t = x_{1t} - x_{2t}$.

To summarize, the null hypothesis considered by Bai and Ng (2006) is equivalent to the definition of DOFs without measurement errors. Hence,it is less general the definition of IOFs considered here.

2.3.2.2 Identifying the IOFs

The idea for identifying the IOFs is similar to the identification of DOFs. If the number of IOFs: m is a priori known, one can use the method in Section 2.3 to select the m out N observed variables that yield the smallest RSS, where the probability of correctly selecting the m IOFs goes to 1 as N and T goes to infinity.

However, when m is not known in practice, one is faced with the choices of both m and $X_{1:m,t}$. To avoid confusion, let m_0 be the true value of m, and let $\hat{m}$ be an estimator of m_0. If $m < m_0$, then any m selected variables cannot span the space of the r factors, otherwise Assumption 15(i) will be violated. Then the sum of RSSs (divided by T) in the regressions of $\tilde{f}_t$ on the selected observed variables will be positive. If $m = m_0$, the sum of RSSs (divided by T) will converge to 0 if $X_{1:m_0,t}$ are selected. However, when $m > m_0$ and $X_{1:m_0,t}$ are among the selected variables, the sum of RSSs (divided by T) will also converge to 0 because adding more regressors never increases the RSSs. To solve this problem, we need to impose some penalty functions to avoid adding extra regressors.

To do so, let us define:

$$[\hat{m}, \hat{n}_1, \hat{n}_2, \ldots, \hat{n}_{\hat{m}}] = \underset{r \le m \le m_{max}, n_1 : n_m}{\arg\min} \Big(S(n_1 : n_m, \hat{A}) + m \cdot p(N,T) \Big), \tag{2.9}$$

where $S(n_1 : n_k, \hat{A})$ is as defined in Section 2.3.1, m_{max} is a predetermined constant, and $p(N,T)$ is a penalty function depending on N and T. The following theorem constitutes the main result of this paper:

Theorem 2.3. *Under Assumptions 11, 12, 13 and 15,*

$$\mathbb{P}[\hat{m} = m_0, (\hat{n}_1, \ldots, \hat{n}_{\hat{m}}) = (1, \ldots, m_0)] \to 1$$

as $N, T \to \infty$, *if* $m_{max} \ge m_0$, $p(N,T) \to 0$ *and* $\delta_{N,T}^2 p(N,T) \to \infty$ *as* $N, T \to \infty$.

The estimation procedure in Section 2.3.1 is repeated for different values of m, and we add a penalty term to the object function. Theorem 2.3 implies that one can identify the number of IOFs and the IOFs simultaneously with probability approaching to 1 as N and T increase.

Since the penalty functions in our procedure and those considered by Bai and Ng (2002) have to satisfy the same conditions, we can use some of their choices that have been proved successful in determining the number of factors. Particularly, we consider the following three penalty functions:

$$p_1(N,T) = \left(\frac{N+T}{NT}\right) \ln\left(\frac{NT}{N+T}\right), \; p_2(N,T) = \left(\frac{N+T}{NT}\right) \ln\left(\delta_{N,T}^2\right), \; p_3(N,T) = \frac{\ln \delta_{N,T}^2}{\delta_{N,T}^2}.$$

These penalty functions have the same asymptotic properties but may perform differently in finite samples (see Bai and Ng, 2002) for a detailed discussion). The finite sample properties of our method using these functions are studied in the Section 2.5.

Remark 1: The above results still hold as long as $f_t - BX_{1:m,t} = o_p(1/\delta_{N,T})$. Compared to the case of DOFs where the measurement error is allowed to be $o_p(1)$, a much smaller measurement error is allowed here. However, in the case that $f_t - BX_{1:m,t} = o_p(1)$, we can still correctly identify the observed factors but m may be over estimated, i.e., some irrelevant variables will also be selected.

Remark 2: The data set used to estimate the factors (X_t) can be different from the data set (Y_t) from which we search for the observed factors. For example, to study the risk factors in financial market, one can use the cross section of assets returns to estimate the factors, and then compare the estimated factors with a panel of macro variables. We can decompose Y_t as the sum of $L(Y_t|f_t)$ and an error term e_t, where $L(Y_t|f_t)$ is the linear projection of Y_t on f_t. Theorem 2.3 still holds if Y_t contains the observed factors and e_t satisfies Assumption 15. Therefore, our method allows the factors to be general functions of some observed variables,

e.g., $f_t = x_{1t} + \theta_1 x_{1,t-1} + \ldots, \theta_p x_{1,t-p}$, or $f_t = x_{1t} + \phi x_{1t}^2$. In those cases, the candidates Y_t should be $[X_t, X_{t-1}, \ldots, X_{t-p}]$ or $[X_t, X_t^2]$.

2.3.2.3 Practical implementation

In the previous discussion, we have assumed that the number of factors (r) is known or correctly estimated. However, in practice, the estimated number of factors using different methods usually differ for the same data set. For example, if one applies the test of Onatski (2009) to the U.S macro data set used in Stock and Watson (2009), 2 factors can be found; but if one uses the 6 different information criteria of Bai and Ng (2002) to the same data, the estimated numbers of factors range from 2 to 6. Actually, it is very rare in practice that the number of factors can be uniquely determined by different methods. Therefore, a discussion on how to implement our methods in practice becomes necessary when the number of factors cannot be correctly specified.

When the estimated number of factors $\hat{r}$ is larger than the true one r, Lemma 2.1 does not hold, so that the above-mentioned methods will fail to identify the IOFs (or DOFs). When $\hat{r} < r$, Lemma 2.1 continues to hold, but our methods will not necessarily identify all of the IOFs. To see this, we first write:

$$\tilde{f}_t = H'BX_{1:m,t} + o_p(1) = AX_{1:m,t} + o_p(1)$$

by Lemma 2.1 and Assumption 15(i), where the matrix $A = H'B$ is $r \times m$. Let A_k be the kth row of A, then $\tilde{f}_{kt} = A_k X_{1:m,t} + o_p(1)$. If we apply our procedure to each of the $\tilde{f}_{kt}$ for $k = 1, \ldots, r$, then $\tilde{f}_{kt}$ can only identify those variables corresponding to the non-zero elements of $\bar{A}_k = \text{plim } A_k$. For example, if $A_1 \xrightarrow{p} (1, 0, \ldots, 0)$, $\tilde{f}_{1t}$ can only identify x_{1t}. However, Theorem 2.3 guarantees that the union of the variables identified by $\tilde{f}_{1t}$ to $\tilde{f}_{rt}$ is equal to the IOFs. The reason is that, since H (also plim$H$) is nonsingular and B has no zero columns (Assumption 15(i)), A (also plim$A$) does not have zero columns.

The previous discussion suggests that we can implement our procedure as follows: Instead of regressing all the estimated factors on the observed variables, we run the regression for each of the estimated factors, starting with the first one: $\tilde{f}_{1t}$. For each $\tilde{f}_{kt}$, define:

$$[\hat{m}_k, \hat{n}_1, \hat{n}_2, \ldots, \hat{n}_{\hat{m}_k}] = \underset{r \le m \le m_{max}, n_1:n_m}{\arg\min} \Big(\frac{1}{T} \sum_{t=1}^{T} (\tilde{f}_{kt} - \hat{A}_k X_{n_1:n_m,t})^2 + m \cdot p(N,T) \Big), \quad (2.10)$$

where $\hat{A}_k$ is the OLS estimator and $p(N,T)$ is as defined above. The key here is when to stop the process. If one stops when $k < r$, the union of the selected variables may be a subset of the IOFs; if one stops when $k > r$, some of the selected variables will not belong to the IOFs. The practitioner can combine the results with some economic theory to judge the appropriateness of the selected variables. If some obvious irrelevant variables are selected

for some large k, one should stop the process and restrict attention to the variables already selected. The main advantage of this procedure is that one can at least identify all of the IOFs, at the cost of identifying some non-IOF variables.

Another practical issue is that the computational cost of our method tend to explode as N, r, m and m_{max} increase. As will be shown in the simulations, when $N = 100$, $r = 2$, $m = 3$ and $m_{max} = 4$, the searching process takes about 1 hour. [3] In practice, N is at least around 100 in most cases, and can be as large as thousands in financial data sets. Since the number of factors r usually ranges from 2 to 8 in many applications, if we were to search in the whole set of variables for those cases, the computational cost could be huge.

To solve this problem, we can restrict our attention to a subset of n variables with $n < N$. Theorems 2.2 and 2.3 should still hold if these n variables contain the observed factors (DOFs or IOFs). In practice, a list of n candidate variables can be selected by prior knowledge and/or economic reasoning. In theory, with large samples, our methods should correctly select the observed factors as long as they are contained in the n variables. However, in practice, the accuracy of our approach with finite samples will depend on n: the smaller n, the less time the computation takes, and the more likely that the observed factors are identified. But a smaller n means that one has to exclude more variables and thus it becomes more likely to miss the IOFs. To reach a balance, we should make n as large as possible whenever the computation cost is affordable. The finite sample performances of our methods when selection is restricted to n variables are studied in Section 2.5.

2.3.3 Identification Based on the Lasso

The methods for identifying the DOF and IOF in the previous sections are shown to consistently select the observed factors, but their high computation costs could make them infeasible in practice if the number of candidate variables are too large. In this section we consider a class of more efficient (in terms of computation cost) estimation procedures called the Lasso that can be used for model selection. We assume that the observed factors are included in a large panel of candidate variables Z_t, which could be different from the data X_t used to estimate the factors.

2.3.3.1 Identification of observed factors using the adaptive Lasso

The *Least Absolute Shrinkage and Selection Estimator* (Lasso), introduced by Tibshirani (1996), is a special case of the more general bridge estimators. In this section, we consider a variant of the Lasso called the adaptive Lasso introduced by Zhou (2006).

[3] The calculations are implemented with Matlab 2009 in a PC with a I5 processor and 8GB of RAM.

Define $\mathcal{M} = \{1, 2, \ldots, m\}$, and suppose $f_t = AZ_{\mathcal{M}t}$ for $t = 1, \ldots, T$, where A is $r \times m$ and $Z_{\mathcal{M}t} = [z_{1t}, z_{2t}, \ldots, z_{mt}]'$ is a $m \times 1$ vector of observed factors, m is a finite constant. Then by Lemma 2.1 we have

$$\tilde{f}_t = H_{NT}AZ_{\mathcal{M}t} + \tilde{V}_t = B_{\mathcal{M}}Z_{\mathcal{M}t} + \tilde{U}_t \tag{2.11}$$

where $B_{\mathcal{M}} = H_0 A$ and $\tilde{U}_t = \tilde{V}_t + (H_{NT} - H_0)f_t$, $\tilde{V}_t = \tilde{f}_t - H_{NT}f_t$, and $H_0 = \text{plim}H_{NT}$. Notice that we write H_{NT} instead of H in Lemma 2.1 to stress the dependence of H on N and T. Now suppose we want to select the observed factors which are included in a large panel of candidate variables $Z_t = [Z'_{\mathcal{M}t} \quad Z'_{\mathcal{P}t}]'$, where $Z_{\mathcal{P}t}$ is a $p \times 1$ vector of irrelevant variables which may be correlated with the observed factors $Z_{\mathcal{M}t}$. From (2.11) we can write

$$\tilde{f}_t = BZ_t + \tilde{U}_t \tag{2.12}$$

where $B = [B_{\mathcal{M}} \quad B_{\mathcal{P}}]$, and $B_{\mathcal{P}}$ denotes a $r \times p$ matrix of zeros. We will focus on a specific estimated factor $\tilde{f}_{kt}$, which can be written as

$$\tilde{f}_{kt} = \beta^{(k)'}_{\mathcal{M}0} Z_{\mathcal{M}t} + \beta^{(k)'}_{\mathcal{P}0} Z_{\mathcal{P}t} + \tilde{u}_{kt} = \beta^{(k)'}_0 Z_t + \tilde{u}_{kt}. \tag{2.13}$$

To simplify the notations, from now on we write β_0, $\beta_{\mathcal{M}0}$ and $\beta_{\mathcal{P}0}$ instead of $\beta^{(k)}_0$, $\beta^{(k)}_{\mathcal{M}0}$ and $\beta^{(k)}_{\mathcal{P}0}$. Therefore (2.13) is simplified as

$$\tilde{f}_{kt} = \beta'_0 Z_t + \tilde{u}_{kt} \tag{2.14}$$

where $\beta_0 = [\beta'_{\mathcal{M}0} \quad \beta'_{\mathcal{P}0}]'$. Also notice that $\beta_{\mathcal{M}0}$ contains zeros when $\tilde{f}_{kt}$ is only associated to a subset of $Z_{\mathcal{M}t}$. We can first show that the each estimated factors identify part of $Z_{\mathcal{M}t}$, and then claim that the union of the identified observed factors by each $\tilde{f}_{kt}$ $(k = 1, \ldots, r)$ is $Z_{\mathcal{M}t}$. However, to avoid carrying on the (k) notations in the proof, it is with out loss of generality to assume that all elements of $\beta_{\mathcal{M}0}$ are nonzero, and thus we can use $\tilde{f}_{kt}$ to identify all the observed factors for a specific k.

The adaptive Lasso estimator $\hat{\beta}^L$ of β_0 using $\tilde{f}_{kt}$ can be defined as follows:

$$\hat{\beta}^L = \underset{\beta \in \mathbb{R}^{(m+p)}}{\arg\min} \sum_{t=1}^{T} (\tilde{f}_{kt} - \beta' Z_t)^2 + \lambda_{NT} \sum_{j=1}^{m+p} w_j |\beta_j|, \tag{2.15}$$

where $w_j = |\tilde{\beta}_j|^{-1}$, and $\tilde{\beta}_j$ is an initial estimator of β_{0j}. The object function in (2.15) is just the usual least square object function plus a penalty on the absolute sum of the coefficients weighted by w. The solution to (2.15) is sparse, in the sense that when the penalty λ_{NT} is large enough, only a small subset of $\tilde{\beta}_L$ is nonzero. Therefore the Adaptive Lasso can be used for model selection, where the selected variables are those with nonzero coefficients. Next, we will show that under some assumptions, regression (2.15) will select the observed factors $Z_{\mathcal{M}t}$, i.e., $\hat{\beta}^L_{\mathcal{M}}$ are non zeros and $\hat{\beta}^L_{\mathcal{P}} = \mathbf{0}$.

Following the literature of the Lasso, we assume the candidate variables Z_t are standardized such that $\sum_{t=1}^{T} z_{jt} = 0$, and $T^{-1}\sum_{t=1}^{T} z_{jt}^2 = 1$ for $j = 1, \ldots, m+p$. We write $a =_s b$ for vectors a and b when the signs of their corresponding elements are the same, i.e., $\text{sign}(a_j) = \text{sign}(b_j)$ for all j, where $\text{sign}(x) = 1(-1)$ if $x > 0 (< 0)$ and $\text{sign}(x) = 0$ if $x = 0$. By construction $\beta_{\mathcal{P}0} =_s \mathbf{0}$, and $\text{sign}(\beta_{0j}) \neq 0$ for $j \in \mathcal{M}$. Define $\Sigma_{\mathcal{M}T} = T^{-1} Z_{\mathcal{M}}' Z_{\mathcal{M}}$, where $Z_{\mathcal{M}}' = [Z_{\mathcal{M}1}, \ldots, Z_{\mathcal{M}T}]$. Let τ_{T1} be the smallest eigenvalue of $\Sigma_{\mathcal{M}T}$, and τ_1 be a constant such that $0 < \tau_1 < \tau_{T1}$ *a.s.* for all T, and $b_1 = \min\{|\beta_{0j}| : j \in \mathcal{M}\}$. The following assumptions are needed for regression (2.15) to consistently select the observed factors,

AL1: $r_{NT}\Big(\max_{1\leq j\leq m+p} |\tilde{\beta}_j - \theta_j|\Big) = O_p(1)$ for some constant θ_j and $r_{NT} \to \infty$, and there exists constants C_1 and C_2 such that

$$\min_{j\in\mathcal{M}} |\theta_j| \geq C_1 > 0, \quad \max_{j\notin\mathcal{M}} |\theta_j| \leq C_2 < \infty.$$

AL2: $\lambda_{NT}/T \to 0$, and $\frac{\lambda_{NT}\delta_{N,T}}{(C_2+1/r_{NT})T} \to \infty$ as $N, T \to \infty$.

AL3: Define $s_{\mathcal{M}} = \Big(|\theta_j|^{-1}\text{sign}(\beta_{0j}), j \in \mathcal{M}\Big)$. For some $\kappa < 1$,

$$\mathbb{P}\Big\{T^{-1}\Big|Z_j' Z_{\mathcal{M}} \Sigma_{\mathcal{M}}^{-1} s_{\mathcal{M}} \theta_j\Big| \geq \kappa \text{ for some } j \notin \mathcal{M}\Big\} \to 0 \text{ as } N, T \to \infty. \tag{2.16}$$

Assumption AL1 states that the initial estimator $\tilde{\beta}_j$ is consistent for θ_j uniformly in j, and the lowest rate of convergence is r_{NT}. But the probability limit $\{\theta_j\}$ are not necessarily equal to the true values $\{\beta_{0j}\}$, the only requirement is that θ_j can not be too small for the relevant variables, and can not be too big for the irrelevant variables. In other words, a reasonable initial estimator that are not necessarily consistent for the true β_0 will suffice for Assumption AL1. Assumption AL2 restricts the size of the tuning parameter λ_{NT}. Remember that $\delta_{N,T} = \min\{\sqrt{N}, \sqrt{T}\}$. In the case where N is about the same size of T, this assumption requires that $\lambda_{NT} = o(T)$ and $\lambda_{NT}/\sqrt{T} \to \infty$, while C_2 can not be too large. Assumption AL3 is an analogue to *adaptive irrepresentable* condition of Huang, Ma and Zhang (2008), restricting the correlations between the relevant and irrelevant variables. This condition is generally not easy to verify in practice, but it is trivially satisfied when $\theta_j = 0$ for $j \notin \mathcal{M}$, i.e., the initial estimator $\tilde{\beta}$ are consistent for the true coefficients of the irrelevant variables. Our assumptions are different from those of Huang et al (2008) mainly in four respects. First, we assume the number of observed factors (relevant variables) doesn't diverge as the number of candidate variables increase. Second, the errors in our true regression equation is $o_p(1)$, thus we don't need to impose the conditions on the tail distributions of the errors. Third, as a consequence of our $o_p(1)$ errors, we don't need to restrict the number of irrelevant variables p. Fourth, the order of λ depends not only on T, but also on N — the number of variables we use to estimate the factors. Under above assumptions, we can prove:

Theorem 2.4. *Under Assumptions 11 to 13 and AL1 to AL3,* $\mathbb{P}[\hat{\beta}^L =_s \beta_0] \to 1$ *as* N *and* T *go to infinity.*

Note that the consistency result in Theorem 2.4 is different from the usual consistency results related to probability limits. In this theorem, we show that with probability approaching to one, the estimated coefficients of the irrelevant variables are set *exactly* to 0. Therefore, by selecting the variables that has non zero estimated coefficients, we can identify the observed factors consistently.

As discussed above, the key Assumption AF3 is trivially satisfied when $\text{plim}\tilde{\beta}_j = \theta_j = 0$ for $i \notin \mathcal{M}$, and such initial estimators can simply be the OLS estimator when $(m+p) < T$ because our errors are $o_p(1)$. However, when $(m+p) \geq T$, the OLS estimator is not feasible, then it is not obvious how to get reasonable initial estimators. A possibility is to use the marginal OLS estimators as proposed by Huang et al (2008), the validity of such estimator is still under study.

Compared to the subset search regression methods in Sections 2.3.1 and 2.3.2, the adaptive Lasso is much faster to compute, but the consistency of the variable selection of the adaptive Lasso comes at the price of more stringent conditions when the number of candidates is possibly larger than the number of observations across time and a reasonable initial estimator of the true coefficients is not available.

The above theorem can easily modified to allow for small measurement errors:

Corallary 1. *Suppose* $f_t = AZ_{\mathcal{M}t} + q_t$ *where* $T^{-1/2}\sqrt{\sum_{t=1}^{T} q_t q_t'} = O_p(\phi_{N,T}^{-1})$, *and* $\phi_{N,T}/\delta_{N,T} \to 0$, *then Theorem 2.4 still holds if*

$$\frac{\lambda_{NT}\phi_{N,T}}{(C_2 + 1/r_{NT})T} \to \infty \text{ as } N, T \to \infty.$$

Therefore we allow some small measurement errors that are larger than the estimation errors of the PC.

2.3.3.2 Computations and others

As discussed above, the sparsity of the selected model, or the estimated number of observed factors in our context, depends crucially on tuning parameter λ. There are some theoretical results how to choose λ but they usually rely on some unknown parameters, see Bickel, Ritov and Tsybakov (2009) for example. In practice, cross-validation (CV) are often used to decide the tuning parameter and the weighting parameters, but the results are far from satisfactory. For example, in the simulation exercises of Huang et al (2008) they generate models with $T = 100$ and $p = 200, m = 15$. The number of selected covariates using CV and adaptive Lasso is at least 2 times larger (from 35 to 60) than the true one. Moreover, the computation cost, although affordable, could be very high if the CV is implemented along several dimensions.

A much more efficient algorithm called *Least Angle Regression* (LARS) is proposed by Efron, Hastie and Johnstone (2004). The LARS can easily calculate the entire solution path of the Lasso or the adaptive Lasso estimators. It give the all the solutions for all values of λ . Therefore, from the path of the solutions, one can see how the covariates are selected in each step. The LARS normally takes only p steps and the computation cost is trivial.

There are some other issues when applying the adaptive Lasso to select the observed factors. First, the adaptive Lasso procedure can be applied to several estimated factors, and the selected observed factors using different factors could contain different elements due to sampling noises. We find that the first estimated factor has higher probability of selecting the observed factors compared to the other ones. Second, as we discussed above, it is possible that an estimated factor is only connected to a subset of the observed factors. In this case the observed factors should combine the selected variables using different estimated factors. Finally, even the selected variables are not the underlying factors, they can serve as good predictors of the factors — the original idea of the Lasso is to find a sparse model with good predictive power. This is important because we can predict the unobserved factors without knowing what they are. The finite sample properties of the adaptive Lasso for identifying observed factors are studied in Section 2.5 using simulations.

2.3.4 Weakly Influential Factors

So far, all the results have been derived under the assumption of *strong factors*, i.e., $N^{-1}\Lambda'\Lambda \xrightarrow{p} \Sigma_\Lambda > 0$. However, in some applications the common factors may only have weak effects on the cross section of variables, or they only affect a small number of variables. For example, in the study of stock market returns, Goyal and Santa-Clara (2006) shows that the idiosyncratic errors are the main source of return volatilities. Moreover, Brown (1989) finds that a one-factor structure is supported by a simulated panel of stock returns even he uses multiple factors to generate the data. Harding (2008) shows that the puzzle is due to the weak effects of the common factors — when the factor loadings are small compared to the idiosyncratic errors, only the first eigenvalue of the covariance matrix diverges as N increases, and the remaining eigenvalues are bounded, in contrast with the prediction of Chamberlain and Rothschild (1983) that all the first r eigenvalues diverges. Therefore, the studying of factor structures in applications such as the stock market returns requires a different asymptotic framework where the factor loadings are allowed to be very small.

Onatski (2012) adopted the assumption that $\Lambda'\Lambda \xrightarrow{p} D > 0$ as an approximation for small factor loadings in finite samples (for example $\lambda_i = O_p(1/\sqrt{N})$). He shows that under this and other assumptions the PC estimator of factors are not consistent for the true factor space. In some cases, the estimated factors could be even orthogonal to the true factor space. Our results will not apply here since they rely on the consistency of PC estimators. However, if both strong and weak factors are present, we can still use the information provided by the strong factors to identify some (if not all) of the true factors.

To see this, we adopt the normalization that $T^{-1}F'F = I_r$, $\Lambda'\Lambda$ is orthogonal, and $D = \text{diag}(d_1, d_2, \ldots, d_r)$ with $d_1 > d_2 > \cdots > d_r$. Define

$$\hat{\beta} = (T^{-1}F'F)^{-1}(T^{-1}F'\tilde{F}) = \frac{1}{T}\sum_{t=1}^{T} f_t\tilde{f}_t'$$

as the OLS estimator of $\tilde{f}_t$ on f_t, where $\hat{\beta}_{\cdot k}$ is the kth column of $\hat{\beta}$. By definition, $\hat{\beta}$ can be viewed as a measure of correlation between the estimated factors and the true factor space. Note that by the normalization $T^{-1}F'F = I_r$, the estimated factors are consistent for the true factor space if $\text{plim}\hat{\beta} = I_r$. If $q(\leq r)$ factors are relatively strong such that d_q is above some threshold value, it is shown by Onatski (2012) that: (i) $\hat{\beta}_{kk} \xrightarrow{p} (1+\gamma_k)^{-1/2}$ for $k = 1, \ldots, q$ and some $\gamma_k > 0$. (ii)$\hat{\beta}_{kk} \xrightarrow{p} 0$ for $k = q+1, \ldots, r$. (iii) $\hat{\beta}_{ks} \xrightarrow{p} 0$ for $k \neq s$[4]. Let R_k^2 denote the the R^2 in the regression of $\tilde{f}_{kt}$ on f_t, and using the above results it is easy to see that:

$$R_k^2 = \hat{\beta}_{\cdot k}'(T^{-1}F'F)\hat{\beta}_{\cdot k} = \sum_{j=1}^{r}\hat{\beta}_{jk}^2 \xrightarrow{p} (1+\gamma_k)^{-1} \text{ for } k = 1, \ldots, q, \tag{2.17}$$

where γ_k can be consistently estimated from the data. Therefore, one can use $(1+\hat{\gamma}_k)^{-1}$ as an indicator of how close is $\tilde{f}_{kt}$ to the true factor space. If $(1+\hat{\gamma}_k)^{-1}$ is very close to 1, it means d_k is very large and the first k normalized factors are very strong. So we can use the first k estimated factors to identify the observed factors. The following two step procedure could be easily implemented to select the observed factors:

(1) Estimate $\text{plim}R_k^2$ from the data for $k = 1, \ldots, r$.

(2) For $\text{plim}R_k^2$ sufficiently close to 1 (say above 0.9), use $\tilde{f}_k$ to select the observed factors: choose the set of observed variables $X_{n_1:n_m,t}$ such that the R^2 in the regression of $\tilde{f}_k$ on $X_{n_1:n_m,t}$ is close to $\text{plim}R_k^2$.

Remark 3: In the second step, we can also use the adaptive Lasso to select the observed factors. This procedure is very similar to the subset search method introduced in the previous section, but we don't provide any consistency results for the above selection procedure. However, we believe it is a practical and useful method when data sets such as the stock returns only contain 1 or 2 relatively strong factors.

Remark 4: This procedure may not be able to select the whole set of observed factors. Because the relatively strong factors, which are normalized to have identity covariance matrix, may be a linear combinations of some observed factors. In other words, it only identifies those observed factors whose linear combinations have strong effects.

Remark 5: Unlike the case of strong factors, it not clear how to choose a penalty function to avoid selecting more observed factors than necessary. We suggest to use t statistics in the regressions to select the those variables that are statistically significant as the observed factors.

[4]Under our standard assumption of strong factors, $d_r \to \infty$ such that $\gamma_i \xrightarrow{p} 0$, therefore the PC estimators are consistent.

2.4 Hypothesis Testing

So far we have assumed the existence of observed factors. Nevertheless, it is possible that the factors cannot be approximated by any observed variables, such as the potential GDP growth rate and the natural rate of unemployment. In such a case, it is necessary to design some tests for the null hypothesis $H_0 : f_t = AX_{1:m,t}$ when some observed factors have been selected by our regression methods. In this section, we propose several test statistics for the H_0 based on both individual and multiple regressions. Notice that the H_0 here covers both DOFs and IOFs, because DOFs can be viewed as a special case of IOFs with A being a $r \times r$ nonsingular matrix. We differentiate these two cases in the estimation because the method for identifying DOFs is simpler, although the method for identifying IOFs includes DOFs as a special case.

The key result underlying our tests is the following lemma proved by Bai (2003):

Lemma 2.5. *Under Assumptions 11 to 13 and Assumption F of Bai (2003),* $\sqrt{N}(\tilde{f}_t - Hf_t) \xrightarrow{d} N(0, \Omega_t)$ *if* $\sqrt{N}/T \to 0$ *as* $N, T \to \infty$*, where* $\Omega_t = V^{-1}Q\Gamma_t Q'V^{-1}$*, and* $\Gamma_t = \lim_{N\to\infty} 1/N \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i\lambda_j' E(e_{it}e_{jt})$.

The matrices V and Q are defined in Appendix A. It follows that:

$$\sqrt{N}(\tilde{f}_{kt} - H_k f_t) \xrightarrow{d} N(0, \sigma_{t,k}^2), \tag{2.18}$$

where H_k is the kth row of H and $\sigma_{t,k}^2 = \Omega_t(k,k)$. Our tests are based on the residuals in the regression of the estimated factors on the selected observed variables. Lemma 2.1 and the null hypothesis imply that $\tilde{f}_t = Hf_t + \tilde{V}_t = BX_{1:m,t} + \tilde{V}_t$, where $\tilde{V}_t = \tilde{f}_t - Hf_t$ and $B = HA$. Let $\hat{B}$ denote the OLS estimator of B, then:

$$\tilde{f}_t = \hat{B}X_{1:m,t} + (B - \hat{B})X_{1:m,t} + \tilde{V}_t = \hat{B}X_{1:m,t} + \hat{V}_t,$$

where $\hat{V}_t = (B-\hat{B})X_{1:m,t} + \tilde{V}_t$. It follows that $\sqrt{N}\hat{V}_t - \sqrt{N}\tilde{V}_t = \sqrt{N}(B-\hat{B})X_{1:m,t}$. Therefore $\sqrt{N}\hat{V}_t$ should converge to the same distribution of $\sqrt{N}\tilde{V}_t$ because $\sqrt{N}(B - \hat{B}) = o_p(1)$. To see this, we can write:

$$\hat{B} - B = \Big(\frac{1}{T}\sum_{t=1}^{T} X_{1:m,t}X_{1:m,t}'\Big)^{-1}\Big(\frac{1}{T}\sum_{t=1}^{T} X_{1:m,t}\tilde{V}_t'\Big).$$

By Assumption 15, $\Big(\frac{1}{T}\sum_{t=1}^{T} X_{1:m,t}X_{1:m,t}'\Big) \xrightarrow{p} \Sigma_{1:m}^x > 0$, and

$$\frac{1}{T}\sum_{t=1}^{T} X_{1:m,t}\tilde{V}_t' = \Lambda_{1:m}\frac{1}{T}\sum_{t=1}^{T} f_t\tilde{V}_t' + \frac{1}{T}\sum_{t=1}^{T} e_{1:m,t}\tilde{V}_t'.$$

By Lemma B1 and B2 of Bai (2003), $\frac{1}{T}\sum_{t=1}^{T} f_t\tilde{V}_t'$ and $\frac{1}{T}\sum_{t=1}^{T} e_{1:m,t}\tilde{V}_t'$ are both $O_p(\delta_{N,T}^{-2})$, whereby it follows that $\sqrt{N}(B-\hat{B}) = O_p(\frac{\sqrt{N}}{\min[N,T]})$, which is $o_p(1)$ under the condition that $\sqrt{N}/T \to 0$. As a result of Lemma 2.5 and the previous analysis, the distribution of the residuals $\hat{V}_t$ in the regressions of $\tilde{f}_t$ on $X_{1:m,t}$ can be derived as follows:

$$N\hat{V}_t'\Omega_t^{-1}\hat{V}_t \xrightarrow{d} \chi_r^2, \tag{2.19}$$

$$N\left(\frac{\hat{v}_{kt}}{\sigma_{t,k}}\right)^2 \xrightarrow{d} \chi_1^2, \tag{2.20}$$

where $\hat{v}_{kt}$ is the kth element of $\hat{V}_t$, i.e., the residuals in the regression of $\tilde{f}_{kt}$ on $X_{1:m,t}$.

Based on these results, we can construct two types of tests. The first type is similar to the $A(j)$ test statistics of Bai and Ng (2006). First, we define:

$$\hat{\rho}_t = N\hat{V}_t'\hat{\Omega}_t^{-1}\hat{V}_t \ , \ \hat{\rho}_{t,k} = N\left(\frac{\hat{v}_{kt}}{\hat{\sigma}_{t,k}}\right)^2, \tag{2.21}$$

and

$$\mathcal{A} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}(\hat{\rho}_t > \Phi_{r,\alpha}) \tag{2.22}$$

$$\mathcal{A}_k = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}(\hat{\rho}_{t,k} > \Phi_{1,\alpha}) \text{ for } k = 1,\dots,r. \tag{2.23}$$

where $\Phi_{r,\alpha}$ and $\Phi_{1,\alpha}$ are two constants such that $\mathbb{P}[\chi_r^2 \geq \Phi_{r,\alpha}] = \mathbb{P}[\chi_1^2 \geq \Phi_{1,\alpha}] = \alpha$, and $\hat{\Omega}_t$ is a consistent estimate of Ω_t[5].

Given the results in (2.19) and (2.20), it can be shown that $E\big(\mathbf{1}(\hat{\rho}_t > \Phi_{r,\alpha})\big) = \mathbb{P}[\hat{\rho}_t > \Phi_{r,\alpha}] \to \alpha$. Then, using the Law of Large Numbers (LLN) we can prove the following result:[6]

Proposition 1. *Under Assumptions 11 to 13 and the hypothesis that* $f_t = AX_{1:m,t}$ *for* $t = 1,\dots,T$, $\mathcal{A} \xrightarrow{p} \alpha$ *and* $\mathcal{A}_k \xrightarrow{p} \alpha$ *for* $k = 1,\dots,r$ *if* $\sqrt{N}/T \to 0$ *as* $N,T \to \infty$ *and* e_{it} *is serially uncorrelated for* $i = 1,\dots,N$.

Notice once more that the $A(j)$ test of Bai and Ng (2006) is based on individual regressions of the observed variables on the estimated factors (regress each of $X_{1:m,t}$ on $\tilde{f}_t$), while we do the opposite here (regress each of $\tilde{f}_t$ on $X_{1:m,t}$). As discussed in Section 2.3.2, the advantage of our procedure is that it allows us to consider more general relations between the factors and observed variables. Moreover, it allows us to construct test statistics not only for the individual regressions, but also for multiple regressions.

For each t, the null hypothesis $f_t = AX_{1:m,t}$ can be tested using statistics in (2.21). However, the test statistics defined in (2.22) and (2.23) are average frequency when the null hypothesis

[5] See Bai and Ng (2006) for discussions on the estimation of Ω_t.

[6] The proof is omitted because given the results in (2.19) and (2.20), it is very similar to the proof of Proposition 1 in Bai and Ng (2006).

is rejected, thus they cannot be used in a strict sense because although their probability limits are derived, their distributions remain unknown. To construct a test statistics with known distribution for the join hypothesis: $f_t = AX_{1:m,t}$ for all $t = 1, \ldots, T$, we can use the standardized sum of $\hat{V}_t$:

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sqrt{N} \hat{V}_t \quad \text{and} \quad \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sqrt{N} \hat{v}_{kt}. \tag{2.24}$$

To derive the distributions of the sums in (2.24), we need to impose the following assumptions:

Assumption 16. *These exists a finite constant M such that:*
(i) $N/T \leq M$ and $\sqrt{N}/T \to 0$ as $N, T \to \infty$.
(ii) $E(e_{it}e_{js}) = 0$ for $t \neq s$; $E(e_{it}^2) = \sigma_{ei}^2$ for all t and $1/N \sum_{i=1}^{N} \sigma_{ei}^2 \leq M$ for all N.
(iii) $E\|1/\sqrt{T} \sum_{t=1}^{T} f_t\| \leq M$ for all T.
(iv) For all s:

$$E\left| \frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \sum_{i=1}^{N} \Big(e_{it}e_{is} - E(e_{it}e_{is}) \Big) \right|^2 \leq M,$$

and

$$E\left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \lambda_i e_{is} \right\|^2 \leq M.$$

(v)

$$E\left\| \frac{1}{T\sqrt{N}} \sum_{s=1}^{T} \sum_{t=1}^{T} \sum_{i=1}^{N} f_s \Big(e_{it}e_{is} - E(e_{it}e_{is}) \Big) \right\| \leq M,$$

$$E\left\| \frac{1}{\sqrt{NT}} \sum_{s=1}^{T} \sum_{i=1}^{N} f_s \lambda_i' e_s \right\| \leq M,$$

and

$$\frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \sum_{i=1}^{N} \lambda_i e_{it} \xrightarrow{d} N(0, \Psi),$$

where $\Psi = \lim_{N,T\to\infty} (1/NT) \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T} \lambda_i \lambda_j' E(e_{it}e_{jt})$.

The first part of the above assumptions allows the number of variables (N) to be comparable or even larger than the number of observations (T), which is not restrictive for most available datasets. The second part requires the idiosyncratic errors to be serially uncorrelated with finite variance, while the third part is generally satisfied for zero-mean factors with some mixing conditions restricting the autocorrelations. These three conditions can be replaced by

$$N/T \to 0 \quad \text{and} \quad \sum_{t=1}^{T} |\gamma_N(s,t)| \leq M \text{ for all } s, \tag{2.25}$$

where $\gamma_N(s,t) = 1/N \sum_{i=1}^{N} E(e_{is}e_{it})$. The above conditions allow the errors to be weakly autocorrelated and the factors to have non-zero means, but require N to be much smaller

than T. The remaining parts of Assumption 16 are generally not restrictive since they involve zero-mean summands. Given that the errors are serially uncorrelated (or asymptotically uncorrelated), those assumptions also require them to have weak cross sectional correlations. Given the above assumptions, we can prove the following theorem:

Theorem 2.6. *Under Assumptions 11, 12, 13, 16 and the hypothesis that $f_t = AX_{1:m,t}$ for $t = 1, \ldots, T$, we have:*

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sqrt{N} \hat{V}_t \xrightarrow{d} N(0, \Xi)$$

where $\Xi = V^{-1} Q \Psi Q' V^{-1}$.

The above theorem allows us to construct the following statistics

$$\mathcal{P} = \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sqrt{N} \hat{V}_t \right)' \hat{\Xi}^{-1} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sqrt{N} \hat{V}_t \right) \tag{2.26}$$

and

$$\mathcal{P}_k = \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sqrt{N} \hat{v}_{kt} \right)^2 \hat{\Xi}^{-1}_{k,k}, \tag{2.27}$$

which converge in distribution to $\chi^2(r)$ and $\chi^2(1)$ respectively, where $\hat{\Xi}$ is a consistent estimator of Ξ. In the most simple case where e_{it} is both serially and cross sectionally uncorrelated and $E(e_{it}) = \sigma_e^2$ for all i and t, or $E(e_{it}^2) = \sigma_{et}^2$ for all i, we can use

$$\hat{\Xi} = V_{NT}^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \lambda_i \lambda_i' \right) \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{e}_{it}^2 \right) V_{NT}^{-1}.$$

When e_{it} is homoskedastic across time, i.e., $E(e_{it}^2) = \sigma_{ei}^2$ for all t, one can use

$$\hat{\Xi} = V_{NT}^{-1} \left(\frac{1}{NT} \sum_{i=1}^{N} \lambda_i \lambda_i' \tilde{e}_{it}^2 \right) V_{NT}^{-1}.$$

When e_{it} is serially uncorrelated but cross sectionally correlated, and $E(e_{it} e_{jt}) = \sigma_{e,ij}$ for all t, then

$$\hat{\Xi} = V_{NT}^{-1} \left(\frac{1}{NT} \sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{i=1}^{N} \lambda_i \lambda_j' \tilde{e}_{it} \tilde{e}_{jt} \right) V_{NT}^{-1}.$$

Unlike the statistics $\mathcal{A}$, the statistics $\mathcal{P}$ has a known limiting distribution and thus can be used for testing the null hypothesis. However, the conditions are more restrictive since the error terms are required to be serially uncorrelated.

Remark 6: Bai and Ng (2006) also proposed some statistics for testing the null hypothesis for a group of observed variables using the theory of canonical correlations, but the limiting

distribution of their tests are known only under very restrictive conditions, e.g., f_t is i.i.d normal (or elliptically) distributed. Our test statistics can also be viewed as tests for a group of observed variables, and the limiting distributions are known under more general conditions.

Remark 7: It is also possible to design a test for the null hypothesis $f_t = AX_{1:m,t}$ under the assumption of weak factors as introduced in Section 2.3.4. Note that under the null, the R^2 in the regression of $\tilde{f}_{kt}$ on $X_{1:m,t}$ is the same as (2.14). The distribution of R_k^2 can be derived if the distribution of $\hat{\beta}_{\cdot k}$ is known. But Onatski (2012) shows that only $\hat{\beta}_{1:q,k}$ have a multivariate normal distribution with mean and covariance matrix that can be consistently estimated from the data, for $k = 1, \ldots, q$. Therefore, we can get the asymptotic distribution of $\sum_{j=1}^{q} \hat{\beta}_{jk}^2$ for $k \leq q$ by sampling from the distribution of $(\hat{\beta}_{1k}, \ldots, \hat{\beta}_{qk})$, and compare the critical values with R_k^2. For the null to be accepted, R_k^2 has to be larger than the 5% quantile of the empirical distribution of $\sum_{j=1}^{q} \hat{\beta}_{jk}^2$. We should keep in mind that $R_k^2 = \sum_{j=1}^{r} \hat{\beta}_{jk}^2 \geq \sum_{j=1}^{q} \hat{\beta}_{jk}^2$, therefore rejecting the null using R_k^2 implies rejecting the null using $\sum_{j=1}^{q} \hat{\beta}_{jk}^2$.

Remark 8: The test statistics allow some small measurement errors. To see this, notice that:

$$\begin{aligned}
\hat{f}_t &= Hf_t + \tilde{V}_t \\
&= Hf_t + HAX_{1:m,t} - HAX_{1:m,t} + \tilde{V}_t \\
&= BX_{1:m,t} + H(f_t - AX_{1:m,t}) + \tilde{V}_t \\
&= \hat{B}X_{1:m,t} + (B - \hat{B})X_{1:m,t} + H(f_t - AX_{1:m,t}) + \tilde{V}_t.
\end{aligned}$$

Define $\eta_t = f_t - AX_{1:m,t}$ as the measurement errors, and $\hat{V}_t = (B - \hat{B})X_{1:m,t} + H(f_t - AX_{1:m,t}) + \tilde{V}_t$ as the residuals in the OLS regressions. It has been shown that if $\sqrt{N}/T \to 0$, $B - \hat{B} = o_p(1/\sqrt{N})$ and $\tilde{V}_t = O_p(1/\sqrt{N})$. Therefore, the proposed test statistics should converge to the same limit distribution as long as $\eta_t = o_p(1/\sqrt{N})$. But of course, even very small measurement errors could contaminate the test statistics in finite samples even they are irrelevant asymptotically.

2.5 Simulations

2.5.1 Directly Observed Factors

In this section, we study the finite sample performance of our method for identifying the DOFs. The following DGP is used: $x_{it} = \lambda_i f_t + e_{it}$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, where f_t are i.i.d multivariate normal vectors with mean 0 and $E(f_t f_t') = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, λ_{ik} and e_{it} are i.i.d random variables drawn from standard normal distributions for $i = r + 1, \ldots, N$,

TABLE 2.1: Probabilities of Correctly Identifying DOFs.

N	T	$\kappa = 0$	$\kappa = \delta_{N,T}^{-2}$	$\kappa = \delta_{N,T}^{-1}$	$\kappa = \delta_{N,T}^{-2/3}$
50	50	1.000	0.980	0.740	0.100
50	100	1.000	1.000	0.870	0.160
50	150	1.000	0.990	0.920	0.210
50	200	1.000	1.000	0.840	0.230
100	50	1.000	1.000	0.950	0.140
100	100	1.000	1.000	1.000	0.600
100	150	1.000	1.000	1.000	0.580
100	200	1.000	1.000	1.000	0.670
150	50	1.000	1.000	0.930	0.100
150	100	1.000	1.000	1.000	0.550
150	150	1.000	1.000	1.000	0.880
150	200	1.000	1.000	1.000	0.930
200	50	1.000	1.000	0.940	0.050
200	100	1.000	1.000	1.000	0.570
200	150	1.000	1.000	1.000	0.820
200	200	1.000	1.000	1.000	0.980

DGP: $x_{it} = \sum_{k=1}^{2} \lambda_{ki} f_{kt} + e_{it}$, where $f_t = (f_{1t}, f_{2t})'$ is multivariate normal with $E(f_{kt}) = 0$, $E(f_{kt}^2) = 1$, and $E(f_{1t} f_{2t}) = 0.5$. $X_{1:2,t} = f_t + \kappa \varepsilon_t$. ε_{jt}, e_{it}, and λ_{ki} are all i.i.d standard normal variables. $\delta_{N,T} = \min[\sqrt{N}, \sqrt{T}]$. The reported numbers are the probabilities of correctly identifying the DOFs: $X_{1:2,t}$ out of 100 replications.

$t = 1, \ldots, T$ and $k = 1, \ldots, r$. Moreover, we let $r = 2$, $\Lambda_{1:2} = I_2$, and the first two variables are generated as $X_{1:2,t} = f_t + \kappa \varepsilon_t$, where ε_{it} are also i.i.d standard normal variables. As has been discussed earlier, the larger the parameter κ, the more difficult is to identify the DOFs.

In the simulations, we report the probability of correctly identifying the DOFs(i.e., the first two variables: $X_{1:2,t}$) out of 1000 replications using the method proposed in Section 2.3, for sample sizes $N, T = 50, 100, 150, 200$, and for 4 different specifications of κ: $\kappa = 0$, $\kappa = \delta_{N,T}^{-2}$, $\kappa = \delta_{N,T}^{-1}$ and $\kappa = \delta_{N,T}^{-2/3}$. Recall that $\delta_{N,T} = \min[\sqrt{N}, \sqrt{T}]$. The results are summarized in Table 2.1.

It can be observed that our method can identify the DOFs correctly with very high probabilities for $\kappa = 0, \delta_{N,T}^{-2}$ and $\delta_{N,T}^{-1}$, even for $N, T = 50$. However, when κ increases to $\delta_{N,T}^{-2/3}$, the probabilities decrease dramatically to less than 30% for $N = 50$ or $T = 50$. Note that $\delta_{N,T}^{-2/3} = 0.27$ when $N = 50$ or $T = 50$, representing a big measurement error. The probabilities increase to more than 50% when $\min[N, T] = 100$ and to more than 80% when $\min[N, T] = 150$.

To study the finite sample properties of the test statistics proposed in Section 2.4 and to compare them to those of Bai and Ng (2006), we generate the simulated data as above except that now κ is fixed to 0. As discussed in Section 2.4, for the DOFs our tests should perform closely to those of Bai and Ng (2006). The simulation results from 1000 replications are summarized in Table 2.2.

Columns 3 to 5 report the averaged statistics defined in (2.22) and (2.23), while columns 6 to 8 display the empirical sizes of the tests defined in (2.26) and (2.27). Finally, the last two

TABLE 2.2: Test with DOFs

N	T	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}$	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}$	$A(1)$	$A(2)$
50	50	0.051	0.058	0.056	0.052	0.047	0.050	0.058	0.059
50	100	0.052	0.054	0.054	0.037	0.054	0.045	0.057	0.057
50	150	0.051	0.054	0.053	0.051	0.041	0.060	0.054	0.056
50	200	0.052	0.053	0.052	0.055	0.059	0.049	0.056	0.056
100	50	0.053	0.057	0.055	0.049	0.053	0.054	0.056	0.056
100	100	0.051	0.054	0.053	0.067	0.056	0.069	0.054	0.054
100	150	0.050	0.053	0.052	0.055	0.047	0.052	0.053	0.053
100	200	0.051	0.053	0.053	0.048	0.043	0.045	0.053	0.054
150	50	0.048	0.057	0.053	0.050	0.045	0.055	0.054	0.054
150	100	0.050	0.053	0.052	0.047	0.050	0.048	0.052	0.053
150	150	0.049	0.053	0.051	0.054	0.047	0.053	0.052	0.052
150	200	0.050	0.052	0.051	0.049	0.047	0.048	0.052	0.052
200	50	0.049	0.057	0.054	0.050	0.053	0.059	0.053	0.054
200	100	0.051	0.053	0.052	0.050	0.041	0.049	0.052	0.053
200	150	0.051	0.052	0.052	0.053	0.052	0.051	0.052	0.052
200	200	0.050	0.053	0.052	0.046	0.056	0.055	0.052	0.052

Note: The DGPs are the same as in Table 2.1 except that $\kappa = 0$. In Columns 3 to 5 are the averaged values of $\mathcal{A}_k$ from 1000 replications. In Columns 6 to 8 are the empirical sizes of the tests $\mathcal{P}_k$ corresponding to the 5% critical value. In Columns 9 to 10 are the averaged values of the $A(j)$ tests of Bai and Ng (2006).

columns show the $A(j)$ statistics of Bai and Ng (2006). It can be seen that all the reported numbers are close to their limiting values (5%), although the $\mathcal{P}_k$ tests tend to be oversized in small sample sizes.

2.5.2 Indirectly Observed Factors

Now we generate data sets with 2 latent factors and 3 observed factors, i.e., $r = 2$ and $m = 3$. The first latent factor is the difference of the first two observed variables: $f_{1t} = x_{1t} - x_{2t}$, and the second latent factors is equal to the third observed variables: $f_{2t} = x_{3t}$. Therefore we can write:

$$f_t = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} X_{1:3,t}.$$

The other parts of the models are generated as in Section 2.5.1. We use the method described in Section 2.3.2 (Equation 3.7) to identify the IOFs, with $m_{max} = 4$. To reduce the computation cost, we restrict the search to subsets of the variables that contain the IOFs. As discussed in Section 2.4, the less variables we consider, the more likely that the IOFs are identified. The results from 500 replications for $n = 10, 20, 30$ are reported in Table 2.3, which shows the probabilities of correctly identifying both the number of IOFs ($m = 3$) and the IOFs.

Several conclusions can be drawn. First, our method performs well in most cases, with high probabilities (more than 80%) of correct identification. Second, $p_3(N, T)$ performs

TABLE 2.3: Probabilities of Correctly Identifying IOFs

N	T	p_1			p_2			p_3		
		$n=10$	$n=20$	$n=30$	$n=10$	$n=20$	$n=30$	$n=10$	$n=20$	$n=30$
50	50	0.722	0.630	0.522	0.622	0.444	0.346	0.916	0.906	0.844
50	100	0.862	0.808	0.744	0.814	0.752	0.674	0.924	0.902	0.878
50	150	0.898	0.844	0.782	0.874	0.808	0.746	0.946	0.910	0.878
50	200	0.904	0.860	0.856	0.886	0.838	0.818	0.940	0.926	0.912
100	50	0.910	0.878	0.830	0.876	0.826	0.748	0.960	0.952	0.938
100	100	0.976	0.970	0.948	0.968	0.930	0.904	0.994	1.000	0.996
100	150	0.990	0.990	0.982	0.984	0.976	0.964	0.998	0.994	1.000
100	200	0.992	0.992	0.984	0.990	0.990	0.982	0.998	1.000	0.998
150	50	0.956	0.908	0.904	0.940	0.884	0.870	0.978	0.964	0.952
150	100	0.984	0.990	0.988	0.976	0.978	0.974	0.998	1.000	1.000
150	150	0.998	0.990	0.993	0.998	0.998	0.986	1.000	1.000	0.998
150	200	0.998	1.000	0.994	0.996	1.000	0.994	1.000	1.000	1.000
200	50	0.952	0.948	0.920	0.944	0.924	0.898	0.974	0.968	0.954
200	100	0.994	0.984	0.988	0.990	0.984	0.982	0.998	1.000	1.000
200	150	1.000	0.998	1.000	0.998	0.998	0.998	1.000	1.000	1.000
200	200	1.000	0.996	0.998	1.000	0.994	0.998	1.000	1.000	1.000

DGP: $x_{it} = \sum_{k=1}^{2} \lambda_{ki} f_{kt} + e_{it}$, where $f_t = (f_{1t}, f_{2t})'$ is multivariate normal with $E(f_{kt}) = 0$, $E(f_{kt}^2) = 1$, and $E(f_{1t} f_{2t}) = 0.5$. $f_{1t} = x_{1t} - x_{2t}$, $f_{2t} = x_{3t}$, e_{it}, and λ_{ki} are all i.i.d standard normal variables. The reported numbers are the probabilities of correctly identifying the IOFs: $X_{1:3,t}$ out of 500 replications for $n = 10, 20, 30$ and 3 different penalty functions p_1, p_2 and p_3.

best among the three penalty functions we consider. Third, the probabilities of correct identification decrease as we increase the number of variables (n) that include the IOFs, but the reductions are not sharp. For most cases, they are less than 2% when we include 10 extra variables in the searching process.

Next we compare our test statistics proposed in Section 2.4 to those of Bai and Ng (2006). The discussions in Section 2.3.2 implies that for the DGPs considered here, the tests of Bai and Ng (2006) will identify x_{3t} as an observed factor but will reject the null hypothesis for x_{1t} and x_{2t}, while our test should identify all of the three observed factors. The simulation results from 1000 replications are reported in Table 2.4.

It can be seen that our tests (columns 3 to 8) still perform good for the all the sample sizes considered. However, the $A(j)$ tests of Bai and Ng (2006) (columns 9 to 12), based on the regressions of IOFs on the estimated factors, fail to converge to 5% for the first two observed factors because they are not directly observed ($x_{1t} - x_{2t} = f_{1t}$). Their tests can only identify x_{3t} as an observed factor because it directly approximates f_{2t}. Hence, these simulation results confirm the superiority of our test in the case of IOFs.

2.5.3 The Lasso and the Adaptive Lasso

In this section the finite sample performances of the Lasso and adaptive Lasso introduced in Section 2.3.3 is studied using simulations. Compared to the adaptive Lasso, the Lasso simply uses the same penalty function for all the coefficients such that $w_j = 1$ for all j. In the first

TABLE 2.4: Test with IOFs

N	T	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}$	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}$	$A(1)$	$A(2)$	$A(3)$
50	50	0.042	0.048	0.044	0.039	0.057	0.045	0.640	0.705	0.065
50	100	0.044	0.047	0.044	0.045	0.045	0.035	0.644	0.710	0.059
50	150	0.045	0.046	0.045	0.046	0.049	0.049	0.650	0.711	0.059
50	200	0.045	0.046	0.044	0.053	0.043	0.056	0.646	0.707	0.059
100	50	0.043	0.053	0.047	0.047	0.044	0.040	0.752	0.794	0.058
100	100	0.047	0.048	0.045	0.037	0.053	0.038	0.753	0.794	0.055
100	150	0.046	0.048	0.046	0.049	0.042	0.042	0.750	0.803	0.054
100	200	0.047	0.048	0.047	0.056	0.053	0.053	0.755	0.780	0.054
150	50	0.045	0.051	0.048	0.051	0.055	0.048	0.803	0.839	0.055
150	100	0.047	0.050	0.048	0.048	0.053	0.041	0.780	0.845	0.053
150	150	0.047	0.048	0.047	0.056	0.065	0.061	0.780	0.835	0.053
150	200	0.047	0.049	0.047	0.046	0.044	0.046	0.799	0.832	0.052
200	50	0.045	0.052	0.048	0.040	0.050	0.048	0.826	0.857	0.054
200	100	0.046	0.050	0.048	0.043	0.048	0.040	0.827	0.859	0.052
200	150	0.048	0.050	0.048	0.051	0.057	0.043	0.828	0.861	0.052
200	200	0.048	0.050	0.048	0.042	0.042	0.045	0.826	0.855	0.052

Note: The DGPs are the same as in Table 2.3. In Columns 3 to 5 are the averaged values of $\mathcal{A}_k$ from 1000 replications. In Columns 6 to 8 are the empirical sizes of the tests $\mathcal{P}_k$ corresponding to the 5% critical value. In Columns 9 to 11 are the averaged values of the $A(j)$ tests of Bai and Ng (2006).

step, r factors are estimated as the first r principal components, which are then used in the Lasso or adaptive Lasso regression to select the observed factors. To be compatible with the applications, we focus on the sample size: $N = 100, T = 200$, and consider different sizes of the subset candidate variables by letting $n = 5, 10, 30, 50, 80$. We also allow the true factors are approximated by the observed factors with measurement errors: $f_t = AX_{1:m,t} + \kappa\epsilon_t$ with $\kappa = 0, 0.1, 0.2, 0.3$. The other parts of the DGPs are the same as those in the previous simulations.

As discussed in Section 2.3.3, we use the more efficient LARS algorithm to calculate the whole solution paths for both Lasso and adaptive Lasso, we refer the details of this algorithm to Efron et al (2004). The solution paths for each of the estimated factors are calculated. The reported numbers in Tables 2.5 and 2.6 are the probability (averages using 1000 replications) of that the solution paths for all the estimated factors contain the observed factors while assuming the number of observed factors m is known. The following conclusions can be drawn from the simulation results: (1) The accuracy decreases as the number of candidates increases, since with more covariates the irrepresentable condition is more likely to be violated—that is, it is more likely that some other variables are close related to the observed factor and are picked up by the selection procedure. (2) The accuracy of Lasso is trivial compared to adaptive Lasso. This is due to the fact that the irrepresentable condition of Zhou (2006), which is necessary for the consistency of model selection, is too stringent for Lasso. (3) The adaptive Lasso has very high accuracy when there is no measurement errors and $n \leq 50$. As the measurement error increases, the observed factors become more like the other variables

and thus the accuracy decreases, although it is very high for the adaptive Lasso when $\kappa = 0.2$ and $n \leq 30$.

In the case of DOF, the overall performance of the adaptive Lasso is worse than the method introduced in Section 2.3, but it is much more attractive in terms of computation cost so it allows us to consider larger set of candidate variables. In the case of IOF, besides the computational advantage, the adaptive Lasso also allows for larger measurement errors than the estimation method in Section 2.3.2, which is actually also a penalized regression with L_0 norm of the coefficients.

2.6 Applications

In this section, we use the methods proposed in this paper to identify the observed factors in financial and macroeconomic data sets. In Section **??** we show that our regression methods are able to identity the well known Fama-French 3 factors in the cross section of portfolio returns constructed by combining different sizes and values. Secondly, in Section **??** we find that the first 3 estimated factors from a panel of macroeconomic variables, which are usually used in factor augmented regressions and factor based forecastings, can be well approximated by 4 observed variables. Finally, we confirm that the Fama-French 3 factors are among the risk factors in stock returns, even though these 3 factors are not enough to span the whole factor space and there exits some other weak factors.

2.6.1 Factors in Portfolio Returns

In this part, we use our method to identify the underlying factors that determine the excess returns of portfolios. It is well known that the Fam-French (FF henceforce) 3 factors, including Market excess return (Market), Small Minus Big (SMB) and High Minus Low (HML), are good approximates of the unobservable risk factors, in the sense that they can explain a large part of the variances of the returns. The purpose of the application is to see that, given that the FF 3 factors are the observed counterpart of the underlying risk factors, and that the estimated factors using PC are consistent for the underlying factors, if our method can successfully identify these 3 factors among a panel of other observed variables. One the other hand, if our method fails to identify the FF 3 factors, we should question the consistency of the estimated factors, or the validity of the FF 3 factors as approximations of the underlying risk factors.

We use two data sets in this empirical study. The first data set consists of the monthly returns of 100 portfolios formed on Size and Book-to-Market, which can be downloaded from the webpage of Kenneth French together with the FF 3 factors. The second data set consists of 151 monthly macro series taken from Stock and Watson (2002b), including variables such

as industrial production, employment, prices, interest rates, and exchange rates. The macro variables are transformed to achieve stationarity, and the transformation methods for each variable can be found in Stock and Watson (2002b). Both data sets range from 1960 to 1997 (T=444).

We first estimate the factors from the panel of portfolio returns, and then identify the observed factors from the macro data set and FF 3 factors. Beside the FF 3 factors, it is widely believed that asset returns are also commonly affected by some macro fundamentals. The use of the macro data set allows us to find the possible connections between macro variables and financial markets.

Before estimating the factors, an important question is how many factors are there. We use two different methods to determine the number of factors for both data sets. The first one is the information criteria (IC) method of Bai and Ng (2002), which penalizes extra factors in a proper way such that the penalty functions help to choose the right number of factors. The second one is Onatski (2010), which is based on the fact that in a factor model with r factors, only the largest r eigenvalues of the covariance matrix explode as the number of variables go to infinity, while the remaining eigenvalues are bounded. The method of Bai and Ng (2002) is usually criticized for overestimating the number of factors, the method of Onatski (2010) is shown to have better finite sample performance when there are non-trivial cross sectional correlations between the idiosyncratic errors.

The estimation results for the number of factors are reported in Table 2.7. It can be seen that for the panel of portfolio returns, 3 to 5 factors are found using different ICs of Bai and Ng (2002), while Onatskis method identifies 3 factors. For the macro data set, the estimated numbers using ICs are all 10, much larger compared to the number (3) found by Onatski's method. We then split the sample by 1980 (for reasons discussed below) and estimate the number factors for each subsamples. The results from Onatski (2009) is the same for both data sets: 4 factors for samples from 1960 to 1980 and 3 factors from 1980 to 1997. The results from Bai and Ng (2002) are less consistent: for the financial data set, the estimated numbers range from 3 to 7 for the two subsamples, and the estimated numbers from subsamples are usually larger than those from the full sample. For the macro data set, the selected numbers of factors using ICs are almost all 10 for each subsamples.

As discussed in Chen Dolado and Gonzalo (2011), the differences in the numbers of factors between subsamples and full sample usually imply structural breaks in the factor model, e.g., the breaks in the factor loadings or the change of factor numbers. However, the number of factors in the full sample should be no less than the number of factors in the subsamples, if the number of factors are correctly estimated. Therefore, the differences of the estimated factors between subsamples and full sample are more likely due to the estimation errors of the two methods in finite samples. Finally, the results in Table 2.7 strongly favors the specification of 3 factors for both data sets in the full sample. We implement the Wald test proposed in Chen Dolado and Gonzalo (2011) for the stability of factor loadings, and find

strong evidence of structural breaks around 1980 for both data sets. For the results to be robust, in the following study we apply our methods to the full sample and both subsamples before and after 1980.

We first estimate the factors from the panel of returns, and then form a list of 50 candidates for the observed factors from the panel of macro variables and FF 3 factors, based on their correlations with the estimated factors and their economic meanings. As discussed in Section 2.3.2, by creating such a list of candidates, we can significantly reduce the computation cost to an affordable level. Table B.1 in the appendix shows the full list of these 50 candidates including their short names, full names and transformation methods. Besides the FF 3 factors, these 50 candidates include the usual macro variables such as industry production, various interest rates, monetary measures, inflations and consumptions, which have often been considered as the main economic factors that affect the financial market in previous studies.

Finally, we identify the observed factors with each of the estimated factors, starting with the first one, and apply our two type of test statistics to each set of identified observed factors. The results are reported in Table 2.8. For each of the estimated factor, we report the minimized object function in (2.10) with $m = 1, \ldots, 4$ and all the three penalty functions considered in Section 2.3.2. Several interesting results are worth noting: (1) When assuming 3 factors and the existence of DOFs, almost all the 3 estimated factors identify the FF 3 factors as the observed factors, except for the second estimated factor in the second subsample. (2) when we consider the case of IOFs, the first 2 estimated factors identify Market and SMB as observed factors, and the third estimated factors identify HML in addition to Market. (3) If a forth factor is estimated, the observed factors identified by it are mainly interests variables except for the stock market indices, and the minimized values are much higher than those of the first 3 estimated factors, implying the existence of only 3 underlying factors. (4) The results are robust for the whole sample and the two subsamples, which implies that the breaks found in the previous subsection are in the factor dynamics since such breaks will not affect the consistency of the estimated factors.

We also report the two type of test statistics for the null hypothesis of exact observed factors defined in Section 2.4, but almost all the tests strongly reject the null, except for the third estimated factor in the second subsample when FF 3 factors are tested. However, the testing results do not necessarily invalidate our identified observed factors. Because for the case of DOFs, we show that our estimation method can tolerate much larger measurement errors than in the test statistics. Therefore, a rejection of the hypothesis of exact observed factors does not contradict with the identification of the observed factors with measurement errors. This is also an advantage of our method compared to that of Bai and Ng (2006).

To provide a rough estimate of the size of the measurement errors, recall that from Remark 8 the residuals $\hat{u}_t$ in the regression of $\tilde{f}_{kt}$ on $X_{1:m,t}$ can be decomposed as

$$\hat{u}_t = (A - \hat{A})X_{1:m,t} + H(f_t - BX_{1:m,t}) + V_t.$$

Given the stochastic orders of $A - \hat{A}$ and V_t, we can get information about the stochastic order of η_t from the residuals $\hat{u}_t$. Suppose $\hat{u}_t = O_p(N^\alpha)$, then a simple estimator of α can be given as:

$$\hat{\alpha} = \log(T^{-1}\sum_{t=1}^{T}\hat{u}_t^2)/2\log(N)$$

because $T^{-1}\sum_{t=1}^{T}\hat{u}_t^2 = O_p(N^{2\alpha})$. The OLS estimation results of the first 3 estimated factors on the FF 3 factors and the estimated αs for each regression are reported in Table 2.9. It is obvious that the estimated α for the $\hat{f}_{2t}$ and $\hat{f}_{3t}$ are much larger than $-1/2$, but still less than 0, and the estimated α for the $\hat{f}_{1t}$ are close to $-1/2$. Given the sizes of $A - \hat{A}$ and V_t, and the fact that $\hat{f}_{2t}$ and $\hat{f}_{3t}$ put most weights on SMB and HML respectively, it is clear that the factors SMB and HML have larger measurement errors than the Market factor, and these measurement errors cause the tests to reject the null of exact observed factors. However, it should be noted that since all estimated α are less than 0, our estimation method should correctly identify the observed factors despite the measurement errors.

2.6.2 Macroeconomic Factors

As discussed in the introduction, following the work of Stock and Watson (2002a,b) and Bernanke et al. (2005), a large part of empirical studies involving factor models use estimated factors as additional regressors in either forecasting equations or Vector Autoregressions. Ludvigson and Ng (2009) is a recent representative example. A commonly used data set in these studies is a panel of monthly macroeconomic time series starting from 1960, first constructed by Stock and Watson. This data set, which is also used is the previous section, is believed to contain rich information about the aggregate economy. Usually a few factors (3 to 8) are estimated to summarize the information of more than 100 macro series. Given that these estimated factors are consistent for the true factor space, the purpose of this section is to see whether the true factors can be approximated by a few observed variables.

Bai and Ng (2006) use a similar data set to study the same question. Their candidates of observed factors include the FF 3 factors, innovations to consumption, inflation and industrial growth rate, a term premium and a risk premium. However, they don't find any strong relations between the factors and these observed variables. Unlike their approach, our regression method searches among all the possible subsets of a large number of observed variables, and doesn't need to impose a short list of candidates. Therefore, our method is expected to explore more information possibly missed by the testing procedure of Bai and Ng (2006). More importantly, as discussed in Sections 2.3.2 and 2.5.2, our method is able to

identify the IOFs when the usual methods based on the regressions of X_t on $\tilde{f}_t$ (such as the test of Bai and Ng 2006) may fail.

The data set we use in this section consists of 131 monthly macro economics variables, from 1964 to 2008. The full list of these variables can be found in Ludvigson and Ng (2010)[7]. Our empirical study differs from previous researches in three important aspects. First, the macro series are usually seasonally adjusted and differenced to achieve stationarity before being used to estimate the factors. However, since the monthly data contain high frequency components, the estimation of factors may be contaminated. To eliminate these high frequency noises, we apply the band pass filter of Christiano and Fitzgerald (1999) to the adjusted data so that only cycle components between 2 to 8 years are retained. Second, consistent with many other studies on the *great moderation*, we find strong evidence of structural instability in the factor structure before 1980 using the test of Chen, Dolado and Gonzalo (2011), therefore we focus on the post-80 sample. We believe that concentrating on a shorter but more stable period may improve the quality of the estimated factors. It should be noted that the second sample period contain enough observations ($T = 336$) for the factors to be well estimated. Third, the estimated number of factors using ICs for the band passed data are very unreliable because all of them are equal to k_{max} (see Table 2.10). Using the method of Onatski (2010) we get 3 factors for the first subsample and 2 factors for the second subsample. Moreover, our instability test also favors the specification of 2 or 3 factors. Therefore, we decide to focus on the first 3 estimated factors, even though many other studies rely on the results of ICs and choose 6 to 8 factors.

In Figure 2.1, we plot the R^2 in the regressions of each macro variables onto each 3 estimated factors, as well as regressions onto all 3 estimated factors. The usual practice of interpreting the factors is based on these R^2. For example, a estimated factor that highly correlated with the price variables is labelled as an *inflation factor*. However, from Figure 2.1 we cannot give an explicit explanations for each of the estimated factors, since all of them are correlated with more than 1 groups of variables. Moreover, it seems that the third estimated factor is not possible to be approximated by any available observed variables, since the R^2 in this graph rarely exceed 50%. The last graph in Figure 2.1 shows the first 3 estimated factors are able to explain the a large part of variations for most of the series.

The identified observed factors using our regression-based methods up to $m = 3$ are reported in Table 2.11 along with the values of the object functions and the R^2. The details of these selected variables are given in Table 2.12. However, by checking the sample covariance matrix of these variables, we find that the smallest eigenvalues are very close to zero, indicating that these variables may be linearly correlated, and that we can further reduce the number of selected variables. Moreover, some of the variables in the data set are very similar to each other and thus provide virtually the same information. For example, the 6th (IPS10) and 14th (IPS34) variables in the data set are both measures of industrial productions and have a

[7] We thank the authors for providing their data set on their webpages.

correlation of 0.98. Although the latter is chosen by our methods, the former is an aggregate variables and thus is a better option to be the observed factors.

Based on the above reasons, we eliminate from the list those variables that can be represented closely by the others, and replace IPS34 by IPS10 (Total industrial production index). Finally, we get a list of $m = 4$ observed factors. Except IPS10, we also include 1 financial variable (84), 1 inflation variable (114), and 1 interest spread (101). Since $m = 4 > r = 3$, we should form some linear combinations of the 4 observed factors to approximate the 3 underlying factors. We find the following specification of the true factors provides good fits with the estimated factors:

$$f_{1t} = \text{IPS10}, \quad f_{2t} = FSPXE + SFYBAAC, \quad f_{3t} = PUNEW, \tag{2.28}$$

where the full names of these variables are given in Table 2.12. Therefore, from the 3 estimated factors, we identify the true factors which are simply linear combinations of some observed variables. The first factor is a *real factor* because it is the industrial production growth rate; the second factor is the sum of stock price index and a yield spread, so it can be called the *financial factor*: the third factor can be labelled as the *inflation factor* since it is simply the inflation rate measured by CPI.

To see how well the factors are approximated by our selected observed factors, we run the following 2 regressions:

1. (Unrestricted regression) Regress $\tilde{f}_t$ on all 4 selected observed factors;
2. (Restricted regression 1) Regress $\tilde{f}_t$ on f_t defined in (2.28);

The fitted values of the above regressions and $\tilde{f}_t$ are plotted in Figure 2.2[8].

First, we can see that the the fitted values are very close to $\tilde{f}_t$, indicating that the space of the true factors are well spanned by the observed factors we found. Second, the fitted values of the restricted regressions and unrestricted regressions are almost indistinguishable from each other, confirming that our specification for f_2 is correct.

2.6.3 Factors in Stock Market

Unlike the returns of portfolios studied in Section 2.6.1, the variances of stock returns are dominated by idiosyncratic errors, making it difficult to identify all the observed factors for reasons discussed in Section 2.3.4. The purpose of this section to check whether the FF 3 factors will be selected by the relative strong factors using our regression based method.

Bai and Ng (2006) use their test statistics to study the same problem, but their tests only identify the Market factor, while the SMB and HML factors are rejected as the observed

[8]The F-test for the imposed restriction gives a p value of 0.758.

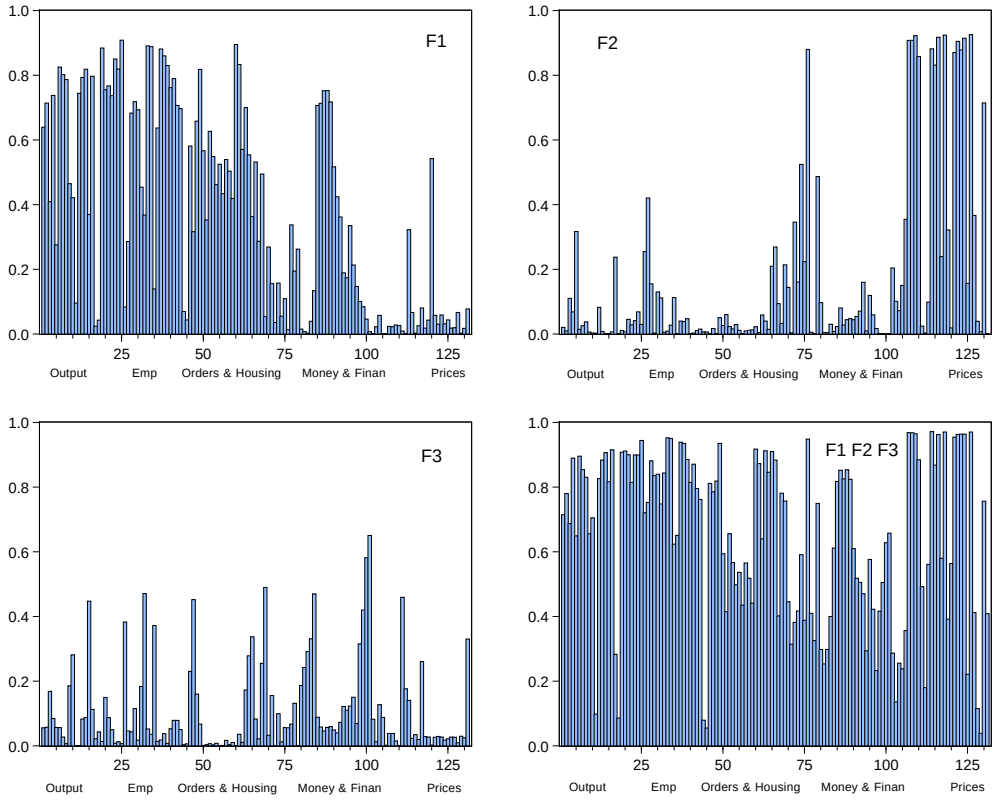


FIGURE 2.1: R^2 in the regressions of the variables listed on the x axis onto the estimated factors.

factors even though they have relatively high correlations with the estimated factors. Using a test procedure as discussed in Remark 7, Onatski (2012) also rejects the null that the common factor space in the stock returns are spanned by the FF 3 factors. The results of Bai and Ng (2006) can be partly explained by the presence of weak factors, which make the some of estimated factors inconsistent. Onatski (2012) explicitly takes into account the possible inconsistency of PC estimators, but the asymptotic distribution of his test is derived under the exact relationship $f_t = BX_{1:m,t}$ for all t, which is too restrictive since we have shown in Section 2.6.1 the measurement errors of SMB and HML are large enough to reject the null.

In this section, we will use the two step procedure proposed in Section 2.3.4 to identify the observed factors. The data set we use is a panel of 747 monthly stock returns from 1980 to 2008. We take all the stocks that are available for the whole sample period from CRSP. The candidate variables from which we search for the observed factors is the monthly macro data set used in the previous section plus the FF 3 factors. The estimates of q is 2 using the method of Onatski (2010), implying 2 relatively strong factors. We then use the method of Onatski (2012) to calculate the limit of R^2 defined in (2.14) for the first 2 estimated factors. The results are $\text{plim}R_1^2 = 0.9971$, $\text{plim}R_2^2 = 0.9591$. Both numbers are quite close to 1, so we

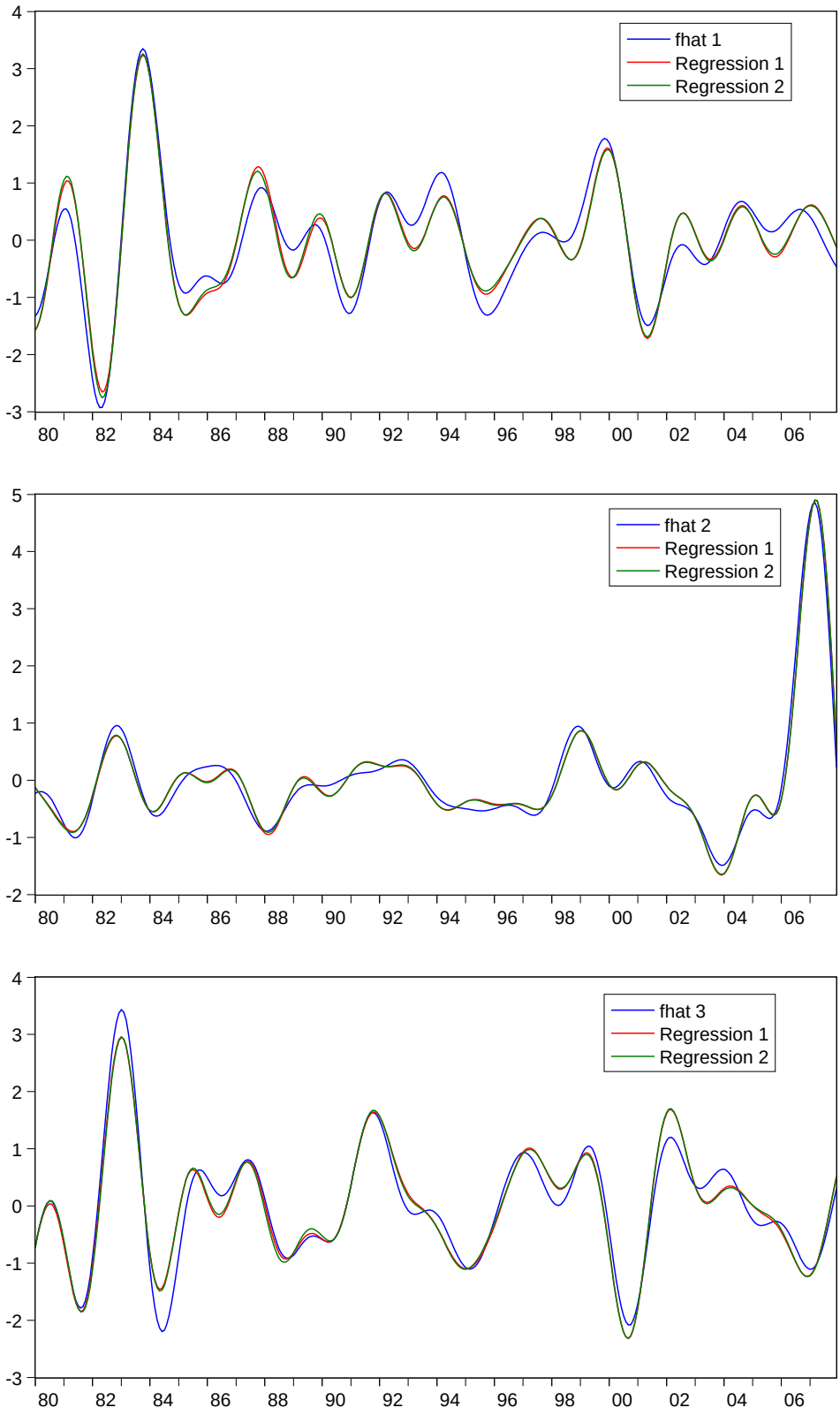


FIGURE 2.2: The estimated factors and the fitted values in the regressions 1 and 2.

can use the first 2 estimated factors for identification. The results using our subset-search methods are reported in Tables 2.13 and 2.14 for $\tilde{f}_{1t}$ and $\tilde{f}_{2t}$.

It can be seen that $\tilde{f}_{1t}$ identifies the FF 3 factors with a R^2 around 90%, and $\tilde{f}_{2t}$ also selects the 3 factors with other variables such as price and interest rate, but with a much lower R^2 (less than 50%). Given the almost consistency of the first estimated factor ($\text{plim}R_1^2 = 0.9971$), and the fact that the FF 3 factors are measured with nontrivial errors, a R^2 as high as 90% is a solid evidence that the FF 3 factors are the observed factors. However, the low R^2 compared to $\text{plim}R_2^2 = 0.9591$ for $\tilde{f}_{2t}$ suggests that those factors we found are not enough to span to whole space of the true factors.

2.7 Conclusion

In this paper, we have studied the identification of the factors in large dimensional FM. The observed variables that can span the space of the true factors are called observed factors. To identify these observed factors and thus provide interpretations to the orthogonal factors estimated by the method of PC, the estimated factors are regressed on some subsets of the observed variables, and the identified observed factors are those which minimize the RSS in the regressions. We show that, if the observed factors exist, this procedure should identify them with probability approaching 1 as N and T go to infinity. We also prove the adaptive Lasso is able to consistently identify the observed factors with much lower computation cost but under more stringent conditions. The problem of how to construct a reasonable initial estimator in the adaptive Lasso is in our research agenda. To test the the assumption that the selected observed factors are indeed observed factors, we propose some test statistics based on individual regressions as well as multiple regressions. We show that our test statistics are more general than those of Bai and Ng (2006). But since all these tests are designed for a exact relationship between the factors and observed variables, the null hypothesis are often rejected in practice. Constructing a test statistics that allows for measurement errors is also an interesting and challenging problem.

TABLE 2.5: Identifying DOF using Lasso and adaptive Lasso.

	$n=5$		$n=10$		$n=30$		$n=50$		$n=80$	
κ	Lasso	A. Lasso	Lasso	A. Lasso	Lasso	A. Lasso	Lasso	A. Lasso	Lasso	A. Lasso
0	0.016	1.000	0.000	1.000	0.000	0.998	0.000	0.996	0.000	0.412
0.1	0.016	1.000	0.000	1.000	0.000	0.975	0.000	0.793	0.000	0.033
0.2	0.018	0.972	0.000	0.900	0.000	0.451	0.000	0.065	0.000	0.000
0.3	0.020	0.791	0.000	0.470	0.000	0.032	0.000	0.000	0.000	0.000

Note: Probabilities of correctly selecting the observed factors by the solution paths of Lasso and adaptive Lasso using all the estimated factors. The DGP is the same as Table 2.1 for $N = 100, T = 200$. The constant κ controls the size of measurement errors and takes values in $[0, 0.1, 0.2, 0.3]$. n is the number of candidate variables to be considered.

TABLE 2.6: Identifying IOF using Lasso and adaptive Lasso.

	$n=5$		$n=10$		$n=30$		$n=50$		$n=80$	
error	Lasso	A. Lasso	Lasso	A. Lasso	Lasso	A. Lasso	Lasso	A. Lasso	Lasso	A. Lasso
0	0.023	0.950	0.000	0.954	0.000	0.947	0.000	0.904	0.000	0.299
0.1	0.011	0.959	0.000	0.947	0.000	0.882	0.000	0.599	0.000	0.022
0.2	0.020	0.913	0.000	0.758	0.000	0.259	0.000	0.037	0.000	0.000
0.3	0.008	0.724	0.000	0.375	0.000	0.021	0.000	0.000	0.000	0.000

Note: Probabilities of correctly selecting the observed factors by the solution paths of Lasso and adaptive Lasso using all the estimated factors. The DGP is the same as Table 2.3 for $N = 100, T = 200$. The constant κ controls the size of measurement errors and takes values in $[0, 0.1, 0.2, 0.3]$. n is the number of candidate variables to be considered.

TABLE 2.7: The estimated number of factors using the information criteria of Bai and Ng (2002) (PC_i and IC_i) and the method of Onatski (2010), with $r_{max} = 10$.

	Samples	PC_1	PC_2	PC_3	IC_1	IC_2	IC_3	Onatski	T	N
Portfolios	1960-1996	5	4	5	3	3	3	3	444	94
	1960-1980	5	4	6	3	3	4	4	240	94
	1980-1996	5	4	7	4	3	5	3	204	94
Macro	1960-1996	10	10	10	10	10	10	3	444	153
Variables	1960-1980	10	9	10	10	8	10	4	240	153
	1980-1996	10	10	10	10	10	10	3	204	153

TABLE 2.8: Identification of observed factors for the returns of portfolios

	1960 − 1996				1960 − 1980				1980 − 1996			
	$m=1$	$m=2$	$m=3$	$m=4$	$m=1$	$m=2$	$m=3$	$m=4$	$m=1$	$m=2$	$m=3$	$m=4$
$\hat{f}_{1t}$	Market	Market SMB	Market SMB HML	FYGT10 Market SMB HML	Market	Market SMB	Market SMB HML	HSBR Market SMB HML	Market	Market SMB	Market SMB HML	FYGT10 Market SMB HML
p_1	0.1660	**0.1438**	0.1939	0.2497	0.1709	**0.1540**	0.2064	0.2684	0.1685	**0.1641**	0.2273	0.2910
p_2	0.1685	**0.1488**	0.2013	0.2596	0.1758	**0.1637**	0.2211	0.2880	**0.1744**	0.1758	0.2449	0.3145
p_3	0.1583	**0.1283**	0.1706	0.2187	0.1569	**0.1259**	0.1643	0.2123	0.1521	**0.1313**	0.1781	0.2255
$\mathcal{A}$	0.7410	0.3491	0.2117	0.2117	0.7375	0.4125	0.2208	0.2250	0.6912	0.2745	0.1912	0.2010
$\mathcal{P}$	739.4618	81.3144	31.9472	31.9472	613.5771	73.8801	20.4605	19.653	370.3764	27.6731	16.1182	16.0050
$\hat{f}_{2t}$	SMB	Market SMB	Market SMB HML	FSPIN Market SMB HML	SMB	Market SMB	Market SMB HML	FYGT10 FYBAAC Market SMB	SMB	Market SMB	FYGT10 Market SMB	FSPCOM FSPIN Market SMB
p_1	0.5119	**0.2502**	0.2718	0.3243	0.6100	**0.2688**	0.3227	0.3777	0.3651	**0.2611**	0.3101	0.3655
p_2	0.5144	**0.2551**	0.2793	0.3342	0.6149	**0.2786**	0.3373	0.3973	0.3710	**0.2728**	0.3277	0.3891
p_3	0.5042	**0.2347**	0.2486	0.2933	0.5959	**0.2408**	0.2806	0.3215	0.3487	**0.2283**	0.2609	0.3000
$\mathcal{A}$	0.4955	0.2613	0.1937	0.1914	0.4958	0.2417	0.2167	0.4667	0.5735	0.5049	0.4363	0.4260
$\mathcal{P}$	139.1675	48.0188	28.7555	27.3977	156.1892	34.3494	31.8496	82.7160	91.9813	64.1346	53.9793	48.2843
$\hat{f}_{3t}$	HML	Market HML	Market SMB HML	FYGT10 Market SMB HML	HML	FYGT1 HML	Market SMB HML	FYGT1 Market SMB HML	HML	Market HML	Market SMB HML	FYBAAC Market SMB HML
p_1	0.2912	**0.2864**	0.3110	0.3594	**0.2495**	0.2977	0.3318	0.3895	0.3535	**0.3115**	0.3279	0.3779
p_2	0.2937	**0.2914**	0.3184	0.3694	**0.2544**	0.3075	0.3465	0.4091	0.3594	**0.3233**	0.3456	0.4015
p_3	0.2834	**0.2709**	0.2877	0.3284	**0.2355**	0.2696	0.2897	0.3333	0.3371	**0.2788**	0.2788	0.3124
$\mathcal{A}$	0.3446	0.2072	0.1351	0.1351	0.2000	0.1750	0.1167	0.1208	0.4265	0.4265	0.0833	0.0833
$\mathcal{P}$	51.3162	29.7754	16.4461	15.2645	18.5583	16.4663	11.5260	11.0523	76.2243	67.4627	5.2939	5.2040
$\hat{f}_{4t}$	PWFSA	FSPIN FSPCAP	FSPIN PSPCAP PWFSA	FSPIN FSPCAP PSFSA PUNEW	FYGT5	FSPIN FSPCAP	PMNV FSPIN FSPCAP	PMNV FSPIN FSPCAP FYGT5	SMB	FSPIN SMB	FSNCOM FYGT10 SMB	FSNCOM FYGT10 GMDC SMB
p_1	**1.0458**	1.0797	1.1254	1.1721	1.0259	**0.9661**	1.0100	1.0564	1.0121	0.9928	**0.9561**	0.9917
p_2	**1.0483**	1.0847	1.1328	1.1820	1.0308	**0.9759**	1.0247	1.0760	1.0180	1.0046	**0.9738**	1.0152
p_3	**1.0380**	1.0642	1.1021	1.1411	1.0118	**0.9381**	0.9679	1.0002	0.9957	0.9601	**0.9069**	0.9262
$\mathcal{A}$	0.1441	0.1329	0.1306	0.1509	0.1917	0.1708	0.1708	0.1750	0.3284	0.3333	0.3186	0.3235
$\mathcal{P}$	66.7837	66.1321	65.5915	63.9249	76.6534	74.9080	74.6898	73.1556	50.7730	46.0358	40.3980	39.2865

TABLE 2.9: Regressions of estimated factors on observed factors.

		Market	SMB	HML	R^2	$\sum \hat{u}_t^2$	$\hat{\alpha}$
60- 96	$\hat{f}_{1t}$	0.1965	0.1239	0.0367	0.9926	3.2460	-0.5364
	$\hat{f}_{2t}$	0.1195	-0.3207	-0.1093	0.8444	68.8791	-0.2032
	$\hat{f}_{3t}$	0.0915	-0.0890	0.3787	0.8698	55.3457	-0.2271
60- 80	$\hat{f}_{1t}$	0.1826	0.1222	0.0452	0.9932	1.6118	-0.5456
	$\hat{f}_{2t}$	0.1584	-0.2966	-0.0491	0.7654	55.7797	-0.1591
	$\hat{f}_{3t}$	0.0492	-0.0618	0.3892	0.8577	32.8917	-0.2167
80- 96	$\hat{f}_{1t}$	0.2123	0.1201	0.0217	0.9941	1.1168	-0.5679
	$\hat{f}_{2t}$	0.1320	-0.3037	0.1852	0.8430	31.1217	-0.2016
	$\hat{f}_{3t}$	0.0381	0.2416	0.3677	0.9201	15.8113	-0.2789

TABLE 2.10: The estimated number of factors using the information criteria of Bai and Ng (2002) and the method of Onatski (2010), with $r_{max} = 10$, for band pass filtered macro data sets from 1964 to 2008.

	Samples	PC_1	PC_2	PC_3	IC_1	IC_2	IC_3	Onatski	T	N
Macro	1964-2008	10	10	10	10	10	10	1	528	131
Variables	1964-1980	10	10	10	10	10	10	3	192	131
	1980-2008	10	10	10	10	10	10	2	336	131

TABLE 2.11: Identification of observed factors for the returns of portfolios

	$\tilde{f}_{1t}$			$\tilde{f}_{2t}$			$\tilde{f}_{3t}$		
	$m=1$	$m=2$	$m=3$	$m=1$	$m=2$	$m=3$	$m=1$	$m=2$	$m=3$
	LUHR	IPS34	IPS34	GMDCN	PMEMP	HSSOU	SFYBAAC	FSPXE	A1M008
		CES002	CES011		PUNEW	PMNV		SFYBAAC	FSPXE
			A0M001			PUC			SFYBAAC
p_1	0.1392	**0.1219**	0.1566	**0.1205**	0.1264	0.1581	0.3982	**0.1878**	0.1971
p_2	0.1427	0.1289	**0.1167**	**0.1240**	0.1334	0.1686	0.4017	**0.1948**	0.2076
p_3	0.1282	**0.0999**	0.1236	0.1095	**0.1044**	0.1251	0.3871	0.1675	**0.1641**
$\bar{R}^2$	0.9090	0.9745	0.9881	0.9277	0.9701	0.9866	0.6501	0.9087	0.9476
$\bar{R}^2$	0.9090	0.9744	0.9880	0.9277	0.9700	0.9865	0.6501	0.9084	0.9473

TABLE 2.12: Details of the selected observed factors

Number	Code	T code	Description
6	IPS10	Δln	Industrial Production Index - Total Index
14	IPS34	Δln	Industrial Production Index - Durable Goods Materials
25	LHUR	Δlv	Unemployment Rate: All Workers, 16 years old and over
33	CES002	Δln	Employees on Nonfarm payrolls - Total Private
36	CES011	Δln	Employees on Nonfarm payrolls - Construction
48	A0M001	lv	Average Weekly Hours
49	PMEMP	lv	NAPM Employment Index
53	HSSOU	ln	Housing Starts - South
63	PMNV	lv	NAPM Inventories Index
64	A1M008	Δln	Mfrs' New Orders, Consumer Goods and Materials
84	FSPXE	Δln	S&P's Composite Common Stock: Price-Earning Ratios
101	SFYBAAC	lv	Baa-Federal Fund Rate
114	PUNEW	$\Delta^2 ln$	CPI - All Items
118	PUC	$\Delta^2 ln$	CPI - Commodities
126	GMDCN	$\Delta^2 ln$	Implicit Price Deflator : Nondurables

TABLE 2.13: Selected observed factors for the stock returns using $\tilde{f}_{1t}$

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
Selected Variables	132	132,133	132,133,134	78,132,133,134,	72,78,132,133,134
p_1	0.2034	0.1922	**0.1637**	0.1831	0.2033
p_2	0.2050	0.1954	**0.1685**	0.1895	0.2133
p_3	0.1972	0.1798	**0.1452**	0.1584	0.1724
$\bar{R}^2$	0.8117	0.8480	0.9024	0.9067	0.9102
$\bar{R}^2$	0.8117	0.8480	0.9018	0.9059	0.9091

Codes for variables: 72: Money Stock; 78: Loans; 132: Market; 133: SMB; 134: HML.

TABLE 2.14: Selected observed factors for the stock returns using $\tilde{f}_{2t}$

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
Selected Variables	133	92,133	91,133,134	91,132,133,134,	91,110,132,133,134
p_1	0.7752	0.6916	0.6320	0.6128	**0.6102**
p_2	0.7768	0.6948	0.6368	0.6192	**0.6182**
p_3	0.7690	0.6792	0.6135	0.5881	**0.5793**
R^2	0.2443	0.3520	0.4355	0.4784	0.5046
$\bar{R}^2$	0.2443	0.3500	0.4321	0.4737	0.4987

Codes for variables: 91: Long run interest rate; 92: Bond yield; 110: Producer price index; 132: Market; 133: SMB; 134: HML.

Chapter 3

The Source of the Aggregate Volatility in Industrial Productions

3.1 Introduction

In most micro-founded macroeconomic models with representative agents, a set of aggregate shocks– the most prominent being an aggregate productivity shock – are assumed to characterize the randomness of the economy. By contrast, another strand of the literature, starting with Long and Plosser (1983), assume the existence of multiple production sectors and sector-specific productivity shocks. These two modelling approaches lead to two different explanations for the source of the aggregate volatility. While the first type of models, with the classical example of Kydland and Prescott (1982) in mind, usually have equilibrium solutions where the dynamics of aggregate variables are driven by aggregate shocks, in the second type of models there has been a long debate on whether sector-specific shocks can generate the observed volatilities of aggregates such as GDP and industrial production (IP). In this paper, we contribute to shedding some light on these issues using recently developed techniques on dynamic factor models (DFM).

Long and Plosser (1983) is the first paper that builds and solves a multi-sector real business cycle (RBC) model, in which N industries produce N different consumption goods, and each sector uses the outputs of other sectors as inputs. Instead of having an aggregate productivity shock, they assume the productivity of different sectors are affected by different sectoral shocks. The solution of their model implies that, in equilibrium, the IP growth rates of different sectors are correlated through input-output linkages even the sectoral shocks are assumed to be mutually independent. As a result of such correlations, the growth rate of the aggregate IP– a weighted average of the sectoral IP growth rates – will not necessarily degenerate to a constant as implied by the law of large numbers (LLN). Thus, a key question to be addressed is: Are these correlations strong enough to entail a positive variance for the aggregate IP growth rates? If the answer is no, then the assumption of sector-specific

shocks (i.e., independent sectoral shocks) should be questioned, and some aggregate shocks that affect all (or most of) the sectors must exist. If the answer is yes, a further question is whether these sector-specific shocks alone can generate the observed aggregate volatility of IP.

Several studies have tried to answer the above-mentioned questions using different approaches. So, Long and Plosser (1987) use a simple factor analysis to study the innovations from a vector autoregressive (VAR) model for the output of 13 manufacturing industries. They find that these innovations, which are proxies for the sectoral shocks, share 1 or 2 common factors and thus that the correlations between the different sectoral outputs are mainly due to the common factors (or aggregate shocks). However, the role of the sector-specific shocks – the idiosyncratic components of the sectoral shocks after extracting the common factors – is not studied. Horvath (1998,2000) builds similar models to that of Long and Plosser (1983) and show that, through input-output linkages, mutually independent sectoral shocks can explain up to 80% of the aggregate volatilities. By contrast, Dupor (1999) uses a similar model to that of Horvath (1998), and claims that the independent sectoral shocks are a poor source of aggregate variability in this model. However, Dupor's results are based on the unrealistic assumption that all sectors are similar as providers of inputs. Forni and Reichlin (1998) study the dynamics of 450 manufacturing sectors in the U.S using DFM and find 2 dynamic factors that account at least 50% of the aggregate output dynamics. Yet, they do not consider how these 2 factors are connected with aggregate and sector-specific shocks, and the relative importance of these two types of shocks. Having realized that the common factors of sectoral output growth rates do not only capture the effects of aggregate shocks but also the effects of sector-specific shocks that propagate through input-output linkages, Foerster, Sarte and Watson (FSW) (2011) calibrate a model which generalizes those of Long and Plosser (1983) and Horvath (1998). Their main finding is that, when compared to the actual data, independent sectoral shocks alone produce noticeably less comovements across sectors, and that two aggregate shocks are needed to replicate the observed volatility in the aggregate output.[1]

In this paper, we study the volatility of the aggregate IP in the U.S through the cross-sectional structure of the disaggregate sectoral IP in 117 four-digit sectors. In particular, we claim that, if the sectoral IP growth rates admit a DFM structure, the growth rate of the aggregate IP will mainly be affected by the common dynamic factors. Using three structural models mentioned above, we show when and how the independent sectoral shocks can propagate into the common factors through input-output linkages. Moreover, using available data sets, we estimate the common factors of the sectoral IP growth rates, and use a penalized regression methods to identify the key sectors whose sector-specific shocks are associated with the factors. The key insight of our theoretical results is that, in contrast to the *granular effects* proposed by Gabaix (2011), the aggregate volatility is mainly affected by the sectors

[1]Other interesting studies on the origins of aggregate volatilities include Gabaix (2011), Burlon (2011), and Acemoglu, Carvalho, Ozdaglar and Tahbaz-Salehi (2012).

that provide inputs for many other sectors, but not necessarily the sectors with the highest weights in the economy. In the empirical study, although we find the common factor is highly correlated with a aggregate shock, the results also provide evidence supporting the propagation mechanism of a sector-specific shock.

The rest of the paper is organized as follows. In Section 3.2, we briefly introduce the three structural models discussed above, and provide theoretical results on the conditions and implications of a DFM structure for the sectoral IP growth rates. In Section 3.3, we first compare the strengths of the three structural models' input-output linkages to propagate sector-specific shocks; we then implement a factor analysis on the sectoral IP growth rates to identify the common factors, which is the main source of aggregate volatility as argued above. Section 3.4 concludes.

3.2 Sector-Specific Shocks, Input-Output Linkages, and Dynamic Factor Models

In this section, we show how independent sector-specific productivity shocks can propagate as common shocks for the sectoral IP growth rates through input-output linkages.

3.2.1 Notations and structural models

Let $X_t = [X_{1t}, \ldots, X_{Nt}]'$ denote the vector of IP growth rates for N different industries at time t, and $A_t = [A_{1t}, \ldots, A_{Nt}]'$ is defined as the vector of *productivity indices* at time t. We assume $\ln(A_t) = \ln(A_{t-1}) + u_t$, where $u_t = [u_{1t}, \ldots, u_{Nt}]'$ is the vector of sectoral shocks.

The vector X_t is said to have a DFM structure if it can be writen as:

$$X_t = B(L)f_t + e_t, \tag{3.1}$$

where f_t is a $q \times 1$ vector of common factors, $B(L) = [b_1(L), \ldots, b_N(L)]'$ is the matrix of dynamic factor loadings, and $e_t = [e_{1t}, \ldots, e_{Nt}]'$ is the vector of idiosyncratic errors satisfying $e_{it} \perp e_{j,t-k}$ for all k and $j \neq i$. The DFM has proved quite useful for characterizing the co-movement of many macro variables because the correlations between a large number of time series can be captured by a few common factors (see, *inter alia*, Geweke 1978, Quah and Sargent 1993, and Forni and Reichlin 1998).

The growth rate of the aggregate IP G_t is a weighted average of the sectoral IP growth rates: $G_t = w'X_t$, where the weighting vector $w = [w_1, \ldots, w_N]'$ is determined by the weight of each sectoral output in total IP. The most important implication of the above DFM

structure is that the volatility of G_t is mainly determined by the common factors f_t, since the idiosyncratic shocks can be diversified by applying the LLN. More precisely, we have

$$G_t = w'X_t \xrightarrow{p} \bar{b}(L)f_t \text{ if } w'e_t \xrightarrow{p} 0,$$

where $\bar{b}(L) = \lim_{N\to\infty} w'B(L)$.

As will be shown in the next section, the DFM structure of X_t can be justified by using methods based on principal component analysis (PCA). The main focus of this paper is the identity of f_t, which is usually claimed to originate from some *aggregate shocks*, i.e., shocks that affect all the sectors. In this paper, however, we show that X_t can have a DFM structure even when the only shocks in the economy u_t are sector specific, i.e., $E(u_t u_t') = I_N$. As in Foerster et al. (2011) and Acemoglu et al. (2012), the key insight is that the sector-specific shocks can propagate as common factors through input-output linkages between different sectors. However, we take a different analytical approach to those adopted in these papers, and further analyze how the number of dynamic factors is determined and how the common factors f_t are connected to the sector-specific shocks e_t.

Generally, we consider models whose solutions have the following moving-average representation:

$$X_t = D(L)u_{t-1} + D_0 u_t, \tag{3.2}$$

where $D(L) = \sum_{h=0}^{\infty} D_h L^h$ is a vector-valued function of the lag operator L. Three different multi-sector real business cycle models with above-mentioned representation are compared, including the models of Long and Plosser (1983), Horvath (1998) and Foerster et al (2011). Since the first two models are special cases of the last one, we start by briefly laying out Foerster's (2011) model and its solutions.

Suppose that there are N industrial sectors in the economy indexed by $j = 1, 2, \ldots, N$. The output of sector j at time t, Y_{jt}, is determined according to the following Cobb-Douglas production function:

$$Y_{jt} = A_{jt} K_{jt}^{\alpha_j} \Big(\prod_{i=1}^{N} M_{ijt}^{\gamma_{ij}} \Big) L_{jt}^{\beta_j}, \tag{3.3}$$

where K_{jt}, L_{jt} are capital and labor inputs used by sector j at time t, M_{ijt} are the materials used by sector j at time t from the output produced by sector i, and $\alpha_j + \sum_{i=1}^{N} \gamma_{ij} + \beta_j = 1$. The capital in each sector j evolves according to:

$$K_{jt+1} = I_{jt} + (1-\delta)K_{jt}, \tag{3.4}$$

where $\delta \in [0, 1]$ is the depreciation rate, and the investment in sector j, I_{jt}, is produced using Q_{ijt} amount of output from sector i according to the following technology:

$$I_{jt} = \prod_{i=1}^{N} Q_{ijt}^{\theta_{ij}}, \quad \sum_{i=1}^{N} \theta_{ij} = 1. \tag{3.5}$$

Finally, a representative agent is assumed to maximize the following standard utility function:

$$E_0 \sum_{t=0}^{\infty} \beta^t \sum_{j=1}^{N} \Big(\frac{C_{jt}^{1-\sigma} - 1}{1-\sigma} - \psi L_{jt} \Big) \tag{3.6}$$

subject to the resource constraint:

$$C_{jt} + \sum_{i=1}^{N} M_{jit} + \sum_{i=1}^{N} Q_{jit} = Y_{jt} \text{ for } j = 1, \ldots, N. \tag{3.7}$$

As discussed earlier, the sector-specific productivity shocks, defined as $u_{jt} = \ln(A_{jt}) - \ln(A_{jt-1})$, are the only source of random variability in this economy. The key parameters determining the linkages between different sectors are the input-output matrix $\Gamma = \{\gamma_{ij}\}$, and the capital use matrix $\Theta = \{\theta_{ij}\}$. The rows of Γ and Θ measure the importance of each sector as provider of materials and capital for other sectors, while the columns of Γ and Θ tell us how each sector combines input from other sectors to produce capital and output.

By log linearizing the first-order conditions and the constraints, the steady state of the economy admits a VARMA(1,1) representation:

$$(1 - \Phi L) X_t = (\Pi_0 + \Pi_1 L) u_t, \tag{3.8}$$

where $X_t = (\Delta \ln Y_{1t}, \ldots, \Delta \ln Y_{Nt})'$, and Φ, Π_0, Π_1 are matrices that depend on the parameters of the model. The above solution can be written as (3.2) by letting:

$$D(L) = (1 - \Phi L)^{-1}(\Phi \Pi_0 + \Pi_1) \text{ and } D_0 = \Pi_0. \tag{3.9}$$

In the economy of Long and Plosser (1983), the agent has a log-utility function over consumption and leisure, and there is no role for capital in the production function ($\alpha_j = 0$ for all j). Moreover, the production is assumed to use materials produced one period before such that: $Y_{jt} = A_{jt}\Big(\prod_{i=1}^{N} M_{ijt-1}^{\gamma_{it}} \Big) L_{jt}^{\beta_j}$. The solution of this economy is a special case of (3.8) with $\Phi = \Gamma'$, $\Pi_0 = I$ and $\Pi_1 = 0$, so that it can be also written as (3.2) with

$$D(L) = (1 - \Phi L)^{-1} \Phi \text{ and } D_0 = I. \tag{3.10}$$

Finally, the model of Horvath (1998) abstracts from labor choice, and assumes full depreciation of capital that is sector specific ($\Theta = I$ and $\delta = 1$). As a result, the solution of their model is: $X_t = (I - \Gamma')^{-1} \Lambda X_{t-1} + (I - \Gamma')^{-1} u_t$, where $\Lambda = \text{diag}(\alpha_1, \ldots, \alpha_N)$. It is a special case of (3.2) by defining

$$D(L) = \sum_{j=1}^{\infty} \Big[(1 - \Gamma')^{-1} \Lambda \Big]^j (1 - \Gamma')^{-1} L^{j-1} \text{ and } D_0 = (1 - \Gamma')^{-1}. \tag{3.11}$$

3.2.2 Input-output linkages and dynamic factor models

In the previous subsection, we have presented three multisector models whose solutions are special cases of the general form (3.2). In this subsection, we focus on the general form solution (3.2), and propose conditions on $D(L)$ and D_0 that allow X_t to have a DFM structure. We use $\tilde{H}$ to denote the conjugate transpose of H (we also use H' for the transpose of H when H contains only real elements). Moreover, $\lambda_i(H)$ denotes the ith largest eigenvalue of the Hermitian matrix H.

First, suppose the sector-specific shocks can be partitioned into two group: $u_t = [u_t^{f'}, u_t^{e'}]'$, and $D(L)$ can be accordingly partitioned as $D(L) = [D^f(L)\ D^e(L)]$, where u_t^f is a $q \times 1$ vector, and $D^f(L)$ is a $N \times q$ matrix. Then (3.2) can be rewritten as

$$X_t = D^f(L)u_{t-1}^f + D^e(L)u_{t-1}^e + D_0 u_t. \tag{3.12}$$

The idea is to show that the effects of the first q sector-specific shocks $u_t^f = [u_{1t}, \ldots, u_{qt}]'$ are transmitted to all the elements in X_t through the structure of $D(L)^f$, while the effects of the remaining shocks (u_t^e) are constrained to affect a small number of variables through the structure of $D(L)^e$ (a specific example is provided below).

Suppose further that $D^f(L) = \sum_{h=0}^{\infty} D_h^f L^h$ and $D^e(L) = \sum_{h=0}^{\infty} D_h^e L^h$. Then, we have that:

$$X_t = \sum_{h=0}^{\infty} D_h^f u_{t-h}^f + \sum_{h=0}^{\infty} D_h^e u_{t-h}^e + D_0 u_t, \tag{3.13}$$

where D_h^f are $N \times q$ matrices and D_h^e are $N \times (N-q)$ matrices. Assuming $E(u_t) = 0$, the spectral density matrix[2] of X_t is defined as

$$\mathcal{S}(\omega) = \sum_{k=-\infty}^{\infty} C_X(k) e^{-ih\omega}, \tag{3.14}$$

where $i = \sqrt{-1}$ and $C_X(k) = E(X_t X_{t+k}')$ is the auto-covariance matrix of X_t. The following assumption is made:

Assumption 17. *u_{it} is a white noise with $E(u_{it}^2) = 1$ for $i = 1, \ldots, N$, and $E(u_{it}u_{jt-k}) = 0$ for all $i \neq j$ and k.*

It then follows from (3.13) that:

$$\begin{aligned}\mathcal{S}(\omega) &= \sum_{k=-\infty}^{\infty} E(X_t X_{t+k}')^{-ik\omega} \\ &= \sum_{k=-\infty}^{\infty}\sum_{h=0}^{\infty} D_h^f \tilde{D}_{h+k}^f e^{-ik\omega} + \sum_{k=-\infty}^{\infty}\sum_{h=0}^{\infty} D_h^e \tilde{D}_{h+k}^e e^{-ik\omega} + \sum_{k=0}^{\infty} D_k \tilde{D}_0 e^{ikw} + \sum_{k=0}^{\infty} D_0 \tilde{D}_k e^{-ikw} + D_0 \tilde{D}_0.\end{aligned}$$

[2]For simplicity we omit $1/2\pi$ in the expression but it is irrelevant for our results.

If we define

$$\mathcal{A}(\omega) = \sum_{h=0}^{\infty} D_h^f \big(I_q \ \mathbf{0}_{q\times(N-q)}\big) e^{ih\omega} + D_0 \tag{3.15}$$

and

$$\mathcal{B}(\omega) = \sum_{h=0}^{\infty} D_h^e \big(\mathbf{0}_{(N-q)\times q} \ I_{N-q}\big) e^{ih\omega} + D_0, \tag{3.16}$$

if follows that

$$\mathcal{S}(\omega) = \mathcal{A}(\omega)\tilde{\mathcal{A}}(\omega) + \mathcal{B}(\omega)\tilde{\mathcal{B}}(\omega) - D_0\tilde{D}_0. \tag{3.17}$$

To derive our main theoretical result, we need to impose the following additional assumptions:

Assumption 18. *(i) For* $c_a(\omega) > 0$, $\lambda_q\big[\mathcal{A}(\omega)\tilde{\mathcal{A}}(\omega)\big] \geq c_a(\omega)N^c$ *almost everywhere in* $[-\pi, \pi]$ *for some constant* $c > 0$*; (ii) There exits a constant* M *such that* $\sup_{\omega\in[-\pi,\pi]} \lambda_1\big[\mathcal{B}(\omega)\tilde{\mathcal{B}}(\omega) - D_0\tilde{D}_0\big] \leq M$ *and* $\sup_{\omega\in[-\pi,\pi]} \lambda_{q+1}\big[\mathcal{A}(\omega)\tilde{\mathcal{A}}(\omega)\big] \leq M$.

Based on (3.17) and Assumption 18, the following result can be derived:

Proposition 3.1. *Suppose* (3.13) *holds, then under Assumptions 17 and 18, the first* q *eigenvalues of the spectral density matrix of* X_t *diverge everywhere in* $[\pi, -\pi]$*, and the remaining* $N - q$ *eigenvalues are uniformly bounded as* $N \to \infty$.

Proof. Similar to the proof of Proposition (1) in Forni et al. (2000), we can show that:

$$\lambda_{q+1}\big[\mathcal{S}(\omega)\big] \leq \lambda_{q+1}\big[\mathcal{A}(\omega)\tilde{\mathcal{A}}(\omega)\big] + \lambda_1\big[\mathcal{B}(\omega)\tilde{\mathcal{B}}(\omega) - D_0\tilde{D}_0\big]$$

and

$$\lambda_q\big[\mathcal{S}(\omega)\big] \geq \lambda_q\big[\mathcal{A}(\omega)\tilde{\mathcal{A}}(\omega)\big],$$

then the desired results follow from Assumption 18. □

We call the first q sectors the **key sectors**, in a sense to be explained below. The implication of Proposition 3.1 is that, if the parameters of the models fall in the space where the solutions of these models satisfy Assumption 18, then the IP growth rates of different sectors admit a unique DFM representation as in (3.1) (see Corollary 1 of Forni et al., 2000), even when the sector-specific shocks are independent white noises. More importantly, the common factors in the reduced-form solutions of X_t are associated with the shocks to the key sectors (u_t^f), and the number of dynamic factors is equal to the number of key sectors (q).

3.2.3 Discussion using a simple example

To better illustrate the ideas behind Proposition 1, let us consider the model of Long and Plosser (1983). It has been shown that the solution of their model can be written as[3]:

$$X_t = \Gamma' u_{t-1} + (\Gamma')^2 u_{t-2} + (\Gamma')^3 u_{t-3} + \cdots + u_t. \tag{3.18}$$

Although the elements of the input-output matrix Γ can be empirically pinned down using the available data sets (see next section), for illustrative purpose, we assume for the moment that

$$\Gamma' = \begin{pmatrix} \gamma & 0 & 0 & \cdots & 0 \\ \gamma & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma & 0 & 0 & \cdots & 0 \end{pmatrix}, \tag{3.19}$$

which, by the definition of Γ, implies that: (i) there is one sector (the first one) that provides materials for all sectors ($\gamma_{ij} = 0$ for $i \neq 1$); (ii) the share of output paid to materials by each sector is the same ($\gamma_{1j} = \gamma < 1$ for all j). With such a simplifying assumption, the solution can be rewritten in terms of (3.13) by defining

$$u_t^f = u_{1t}, \quad u_t^e = [u_{2t}, u_{3t}, \ldots, u_{Nt}]', \quad D_0 = I_N$$

and

$$D_h^f = \begin{pmatrix} \gamma^{h+1} \\ \gamma^{h+1} \\ \vdots \\ \gamma^{h+1} \end{pmatrix}, \quad D_h^e = \mathbf{0}_{N\times(N-1)} \text{ for } h = 0, 1, \ldots, \infty.$$

Further, in this specific case it is easy to show that $\mathcal{B}(\omega)\tilde{\mathcal{B}}(\omega) - D_0\tilde{D}_0 = \mathbf{0}_{N\times N}$, so that Assumption 18(ii) trivially holds. Finally, as regards Assumption 18(i), notice that

$$\mathcal{A}(\omega)\tilde{\mathcal{A}}(\omega) = \begin{pmatrix} a(\omega)b(\omega) + a(\omega) + b(\omega) & a(\omega)b(\omega) + b(\omega) & \cdots & a(\omega)b(\omega) + b(\omega) \\ a(\omega)b(\omega) + a(\omega) & a(\omega)b(\omega) & \cdots & a(\omega)b(\omega) \\ \vdots & \vdots & \vdots & \vdots \\ a(\omega)b(\omega) + a(\omega) & a(\omega)b(\omega) & \cdots & a(\omega)b(\omega) \end{pmatrix},$$

where $a(\omega) = \sum_{h=1}^{\infty} \gamma^h e^{ih\omega}$ and $b(\omega) = \sum_{h=1}^{\infty} \gamma^h e^{-ih\omega}$. For each ω, the above matrix has $N-2$ eigenvalues equal to 0, and the remaining 2 eigenvalues are:

$$\frac{N \cdot a(\omega)b(\omega) + a(\omega) + b(\omega) \pm \sqrt{\left[N \cdot a(\omega)b(\omega) + a(\omega) + b(\omega)\right]^2 + 4a(\omega)b(\omega)(N-1)}}{2}. \tag{3.20}$$

[3] Assuming Γ satisfies the necessary condition for Equation (2) to be causal.

When N is sufficiently large, one of them converges to 0 and the other converges to $N \cdot a(\omega)b(\omega)$.

Hence, since

$$a(\omega)b(\omega) = \left(\sum_{h=0}^{\infty} \gamma^h \cos(h\omega)\right)^2 + \left(\sum_{h=0}^{\infty} \gamma^h \sin(h\omega)\right)^2 > 0$$

for any $\omega \in [-\pi, \pi]$, Assumption 18(i) is satisfied with $q = 1$. Therefore, the solution of the multi-sector model of Long and Plosser (1983) under assumption (3.19) implies that the IP growth rates of the N sectors admit a DFM representation with 1 common dynamic factor. Alternatively, we can directly use (3.10) to show that

$$X_t = \Gamma' X_{t-1} + u_t = \begin{pmatrix} \gamma \\ \gamma \\ \vdots \\ \gamma \end{pmatrix} X_{1t-1} + u_t = \begin{pmatrix} \gamma \\ \gamma \\ \vdots \\ \gamma \end{pmatrix} (1 - \gamma L)^{-1} u_{1t-1} + u_t. \tag{3.21}$$

More importantly, by defining $f_t = (1 - \gamma L)^{-1} u_{1t-1}$, this common factor is associated with the sector-specific shock (u_{1t}) to the key sector (i.e., sector 1)

Under (3.19), the intuition for the above results is clear: the shock affecting the key sector, whose output is the only source of materials for all other sectors, will affect the whole economy even when this shock is uncorrelated with the other sector-specific shocks. As a result, the effects of such shocks will not be averaged out when calculating the growth rate of aggregate IP. In other words, the volatility of the aggregate IP is not determined by the sectors with the largest weights, but by the key sectors that provide most inputs for the other sectors.

Moreover, if we assume the technology shocks are not sector specific but share some common shocks, i.e., $u_t = Hg_t + \xi_t$, where g_t is a vector of r common technology shocks and ξ_t is the vector of independent white noises, then (3.21) can be written as:

$$X_t = \Gamma' X_{t-1} + u_t = \begin{pmatrix} \gamma \\ \gamma \\ \vdots \\ \gamma \end{pmatrix} X_{1t-1} + u_t = \begin{pmatrix} \gamma \\ \gamma \\ \vdots \\ \gamma \end{pmatrix} (1 - \gamma L)^{-1} u_{1t-1} + Hg_t + \xi_t, \tag{3.22}$$

which is a standard factor model with r+1 common factors.

It is noteworthy, however, that the actual input-output matrix Γ constructed using available data sets is much more complicated than (3.19): almost all the sectors use inputs from all other sectors, and the number of key sectors is likely to be larger than one. This is precisely why we propose Assumption 18. Under this assumption, Proposition 1 can be used to identify the key sectors and their number based on the input-output matrix Γ and the capital use

matrix Θ, because the $D(L)$ and D_0 matrices in solution form (3.2) are connected with Γ and Θ in each considered model.

3.3 Empirical Analysis of Sectoral IP Growth Rates Using Factor Models

In our empirical analysis, we focus on the quarterly IP growth rates of 117 sectors over the sample 1972-2007. These sectors correspond to four-digit industries as defined in the North American Industry Classification System (NAICS). The data can be downloaded from Mark Watson's webpage, and we refer to Foerster et al. (2011) for the details of this data set.

3.3.1 Can sector-specific shocks generate aggregate volatility?

In this subsection, we investigate the following question: Can sector-specific productivity shocks — defined as mutually independent white noises, like in Assumption 18 — be propagated into the common factors for the sectoral IP growth rates through input-output linkages? If the answer is affirmative, our results in Section 3.2 imply that it can be inferred that the aggregate IP (a weighted average of the sectoral IPs), will mainly be affected by the common factors that originate from the shocks hitting the key sectors. If the sectoral shocks are all affected by some aggregate productivity shocks, the common factors of X_t (and thus the growth rate of the aggregate IP) will surely be affected by these aggregate shocks. However, the purpose of this subsection is to isolate the effects the of the sector-specific shocks.

The answer to the above question depends on the model specification, because the $D(L)$ and D_0 matrix in (3.2), which determine the strengths of the input-output linkages between different sectors, are model-specific. For example, in the model of Long and Plosser (1983; LP below), $D(L)$ and D_0 only depends on the input-output matrix $\Gamma = \{\gamma_{ij}\}$, since they do not consider the role of capital in the production function. As regards the model of Horvath (1998, H below), $D(L)$ and D_0 are functions of Γ and $\Lambda = \text{diag}(\alpha_1, \ldots, \alpha_N)$, since the assumption is that capital is sector specific. Finally, in the model of Foerster et al. (2011, FSW below), because of its generality these two matrices depend on all the parameters of the model in a rather complex way.

Proposition 1 implies that the number of dynamic factors for X_t is equal to the number of diverging eigenvalues of the spectral density matrix $\mathcal{S}(\omega)$. Under Assumption 17, $\mathcal{S}(\omega)$ for the three models can be written as: $\mathcal{S}(\omega) = D(e^{-i\omega})\tilde{D}(e^{-i\omega})$, where

$$D_{LP}(e^{-i\omega}) = (I_N - \Gamma' e^{-i\omega})^{-1},$$

$$D_{Horvath}(e^{-i\omega}) = (I_N - (I - \Gamma')^{-1}\Lambda e^{-i\omega})^{-1}(I_N - \Gamma')^{-1},$$

$$D_{FSW}(e^{-i\omega}) = (I_N - \Phi e^{-i\omega})^{-1}(\Pi_0 + \Pi_1 e^{-i\omega}).$$

The matrices Θ, Λ and Γ are constructed using the BEA's (Bureau of Economic Analysis) use tables and capital flow tables for 1997 (see Foerster et al. 2011 for details). To facilitate comparison, we also follow the calibration of FSW to choose other parameter values in their model. Figures 3.1 to 3.3 display the five largest eigenvalues of the spectral density matrices of X_t implied by these three models.

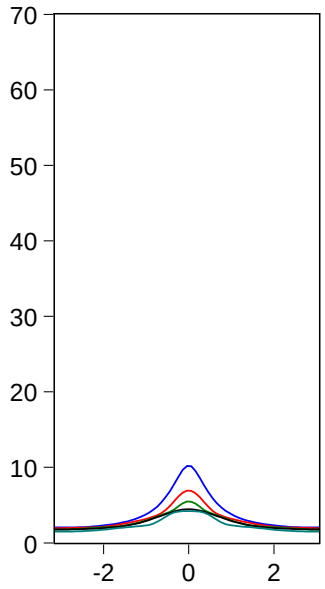


FIGURE 3.1: Long and Plosser

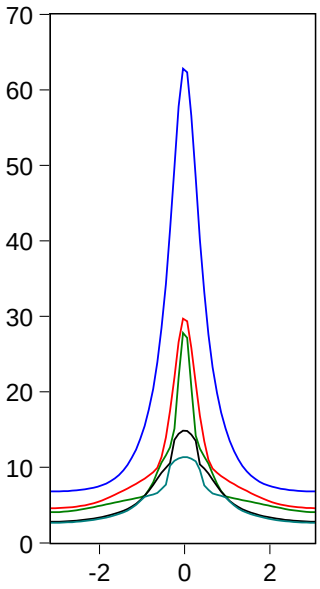


FIGURE 3.2: Horvath

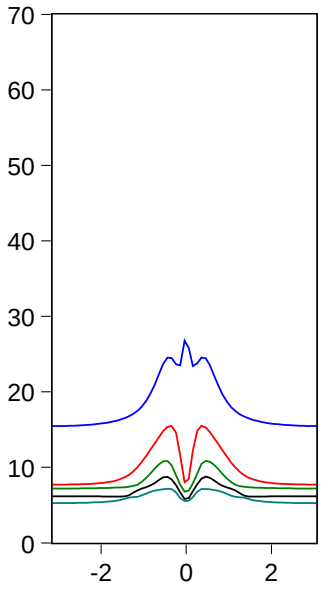


FIGURE 3.3: FSW

The defining feature of a DFM implies that the first q eigenvalues of the spectral density matrix diverges as $N \to \infty$ at each frequency. For a given large enough N, we should therefore observe a clear separation between the first q eigenvalues and the remaining $N - q$ ones. In Figures 3.2 and 3.3 we can see such a separation between the first and the remaining eigenvalues of the corresponding $\mathcal{S}(\omega)$, implying that the propagation mechanism in the H and FSW models is strong enough to imply that even mutually uncorrelated sectoral shocks can be propagated into common factors hitting all sectors. However, in Figure 3.1, there is no clear separation of the eigenvalues in the LP model, on top of being much smaller compared to the eigenvalues in the other two models. It can be concluded that the input-output linkages in the LP model are not strong enough to transmit any of the sector-specific shocks into common factors.

3.3.2 Number of factors and structural breaks

In Section 3.2, we have shown that under Assumption 18, the cross section of sectoral IP growth rates admits a DFM representation: $X_t = B(L)f_t + e_t$. In comparison, we say that X_t

follows a *static factor model* (SFM) if $X_t = LF_t + \varepsilon_t$, where $L = [l_1, \ldots, l_N]'$ is a $N \times r$ matrix of factor loadings, F_t is a $r \times 1$ vector of common factors, and $\varepsilon_t = [\varepsilon_{1t}, \ldots, \varepsilon_{Nt}]'$ is a $N \times 1$ vector of idiosyncratic errors. The SFM is also called as *approximate factor model* when the N is large and the errors are allowed to be cross-sectionally correlated (see Chamberlain and Rothschild, 1983), which is the setting we assume for the sectoral IP growth rates. There are two case where a DFM can be reduced to a SFM. First, when $B(L) = \sum_{h=0}^{p} B_h L^h$ for some finite p, we can write the DFM in terms of SFM by defining $L = [B_0 \quad B_1 \ldots B_p]$ and $F_t = [f_t', f_{t-1}', \ldots, f_{t-p}']'$. Second, when the dynamic loading matrix satisfies $B(L) = B \cdot b(L)$, the DFM can be simplified to a SFM with $L = B$ and $F_t = b(L)f_t$ (this is the case for our simple example in section 3.2.2). In the first case, the number of static factors is associated with the number of dynamic factors by $r = q \times (p + 1)$, while in the second case $r = q$. In other cases, such as the set up considered by Forni et al (2000), it is not possible to convert a DFM to a finite dimensional SFM. Therefore, it is preferable to work with SFM when the number of static factor r is found to be small, because r and the space of F_t can be consistently estimated using methods based on PCA under fairly general conditions (see, e.g., Bai and Ng, 2002).

We first use the testing procedure of Onatski (2009) to estimate the number of dynamic factors for X_t. The test statistic for the hypothesis $q = q_0$ against $q_0 < q \leq q_1$ is constructed as

$$\max_{q_0 < i \leq q_1} \frac{\lambda_i(\omega) - \lambda_{i+1}(\omega)}{\lambda_{i+1}(\omega) - \lambda_{i+2}(\omega)},$$

where $\lambda_i(\omega)$ is the ith largest eigenvalue of the estimated spectral density matrix of X_t at frequency ω. The test can be implemented sequentially to estimate q. Using this procedure, we find estimated numbers of dynamic factors for the sectoral IP growth rates to be 1 for almost all the frequencies between $[-\pi, \pi]$.

Next, to see if X_t can be also represented as a SFM with a small number of static factors, we estimate the number of static factors r using the methods of Bai and Ng (2002), Onatski (2010) and Ahn and Horenstein (2013). The first method uses various information criteria (IC) for choosing the r, and the last two methods rely on the fact that the first r eigenvalues of the covariance matrix of X_t diverges as $N \to \infty$ while the remaining eigenvalues are bounded. Although all these methods provide consistent estimators of r under similar conditions, it has been shown that the Bai and Ng's (2002) method tends to overestimate r especially when the idiosyncratic errors have relatively strong cross-sectional correlations.

The estimated number of static factors for the whole sample period using those methods are reported in the first row of Table 3.1 below. Columns 2 to 7 report the results from using different IC of Bai and Ng (2002). It can be seen that these numbers vary from 2 to 9 depending on which specific IC to use: Notice that the results of IC_1 and IC_2 are more consistent with the results of Onatski (2010) and Ahn and Horenstein (2013), which indicate 2 static factors for the whole sample. To further confirm the specification of 2 static factors for X_t, we also implement the sequential tests of Onatski (2009) for $r_0 = 1, 2$ and

TABLE 3.1: The Estimated Number of Static Factors

Sample	PC_1	PC_2	PC_3	IC_1	IC_2	IC_3	Onatski	A&H
1972 - 2007	4	3	9	2	2	9	2	2
1972 - 1983	6	5	7	2	2	5	7	1
1984 - 2007	3	2	8	1	1	5	1	1

$r_1 = 2, \ldots, 6$. The reported numbers in Table 3.2 are the p values of the test for the null hypothesis $r = r_0$ against $r_0 < r \leq r_1$. Columns 2 and 3 in Table 3.2 imply that the null $r = 1$ is rejected while $r = 2$ can not be rejected for the whole sample. These results provide evidence that X_t can be well characterized by a SFM with only 2 static factors.

TABLE 3.2: Testing the Number of Static Factors

	1972-2007		1972-1983		1984-2007	
	$r_0 = 1$	$r_0 = 2$	$r_0 = 1$	$r_0 = 2$	$r_0 = 1$	$r_0 = 2$
$r_1 = 2$	0.028		0.137		0.200	
$r_1 = 3$	0.050	0.354	0.247	0.354	0.364	0.844
$r_1 = 4$	0.068	0.623	0.292	0.218	0.269	0.201
$r_1 = 5$	0.087	0.775	0.357	0.292	0.329	0.269
$r_1 = 6$	0.103	0.863	0.413	0.357	0.382	0.329

However, as pointed out by Chen, Dolado and Gonzalo (2013; CDG hereafter), the finding of 2 static factors for the whole sample could be the consequence of having one structural break in the factor loadings when there is actually only one static factor. To explore this possibility, we split the whole sample by the end of 1983, a date found to be the beginning of so called *Great Moderation*, and then estimate the number of static factors in each subsample. As can be seen in Table 3.1, the estimated r in the second subsample (1984-2007) is always smaller that that estimated using the whole sample, and in most cases the estimated r is 1. The estimated r for the first subsample (1972-1983) is less consistent, possibly due to the short sample period ($T = 44$) since the consistency of these methods require both N and T to be large. Moreover, the sequential testing procedure of Onatski (2009) can not reject $r = 1$ for both subsamples. We implement the Sup-Wald test developed by CDG (2013) for big structural breaks in the factor loadings, but we cannot reject the null hypothesis of no structural breaks. Since our main purpose is to identify the common factors but not the causes of the Great Moderation, we focus on the second subsample ranging from 1984 to 2007, which is found to be more stable and has a simple structure of 1-factor SFM.

Finally, to further confirm the 1-factor SFM structure for X_t from 1984 to 2007, notice that a unique feature of a approximate factor model with 1 factor is that both the PCA estimates of F_t and a cross-sectional average of X_t can consistently estimate the common factor. More specifically, let $\hat{F}_t$ denote the first principle component of X_t, then according to Theorem 1

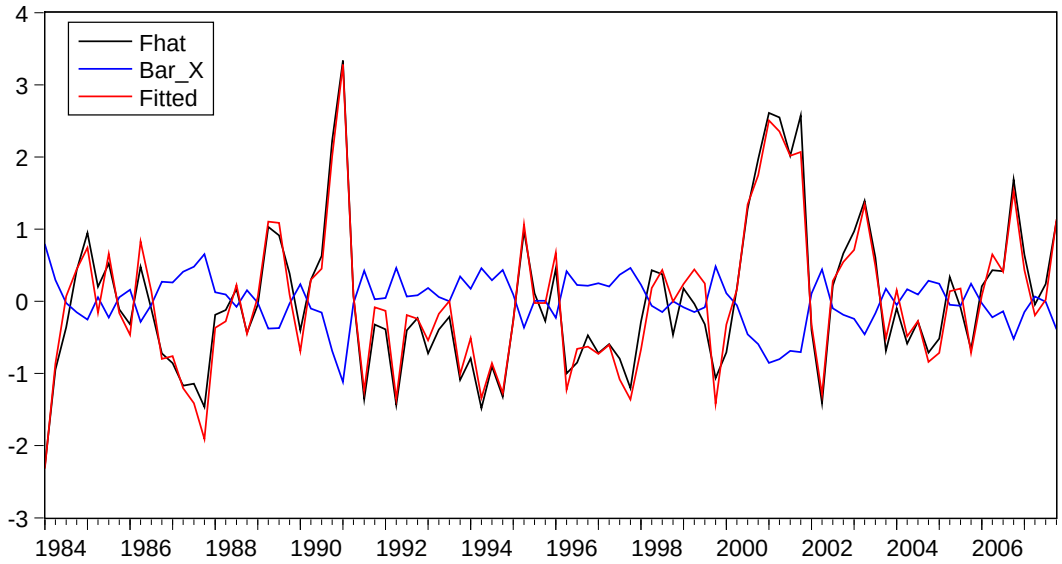


FIGURE 3.4: The Evidence of One Static Factor

of Bai and Ng (2002) we have $\hat{F}_t \xrightarrow{p} cF_t$ for some non-zero constant c; on the other hand, let $\bar{X}_t = 1/N \sum_{i=1}^{N} X_{it}$, it is easy to see that $\bar{X}_t \xrightarrow{p} \bar{L}F_t$ as long as $1/N \sum_{i=1}^{N} \varepsilon_{it} = o_p(1)$, where $\bar{L} = 1/N \sum_{i=1}^{N} l_i$. Therefore, we have $\hat{F}_t = \alpha \bar{X}_t + o_p(1)$ where $\alpha = c/\bar{L}$. In Figure 3.4, we plot $\hat{F}_t$, $\bar{X}_t$ and the fitted values in the OLS regression of $\hat{F}_t$ on $\bar{X}_t$. As can be observed, reassuringly $\hat{F}_t$ and $\bar{X}_t$ are quite close to each other up to a constant, as predicted by a 1-factor SFM.

Another important implication of the 1-factor structure is that the growth rates of the aggregate IP can also be seen as a consistent estimator of the common factor. To see this, notice that the growth rate of the aggregate IP, G_t, can be written as $G_t = w'X_t$ where $w = [w_1, \ldots, w_N]'$ and w_i is the weight of sector i such that $\sum_{i=1}^{N} w_i = 1$. It follows that $G_t = w'X_t \to (w'L)F_t$ as long as $1/N \sum_{i}^{N} w_i \varepsilon_{it} = o_p(1)$. In Chen (2013), G_t is identified as one of the common factors affecting a large panel of macro variables, including consumption, housings, inflation, exchange rate, stock market index, etc. Therefore, the common factor of the sectoral IP growth rates not only affects all the industrial sectors, but also serves as a fundamental shock to the whole economy. In the next subsection, we investigate the nature of this common factor in term of observables.

3.3.3 Identifying the common factor

In this subsection, we try to identify the common factor of the sectoral IP growth rates. First, we have shown in Section 3.3.1 that the common factor of X_t can originate from sector-specific shocks through input-output links; Second, when the sectoral productivity shocks are affected by both aggregate productivity shocks and sector-specific shocks (u_t has

a factor model structure), the common factor of X_t can be associated with the aggregate shocks and some sector-specific shocks. The first key issue to analyze is whether the sectoral shocks are sector specific or they share some common aggregate shocks.

To address this question, notice that, since the sectoral shocks are not directly observable, they can be estimated using the three structural models discussed in Section 3.2. Once the sectoral shocks are estimated as $\hat{u}_t$, we can test whether they have common factors by applying the three methods discussed earlier to estimate the number of factors for $\hat{u}_t$ (r_u).

The results are reported in Table 3.3 below. Again, the estimated number of factors for $\hat{u}_t$ in each model vary a lot, ranging from 0 to 10. However, if we focus on the methods which proved to be successful in estimating r for X_t, e.g., IC_1, IC_2 and *Onatski*, we can see the estimated r_u are 0 for the H and FSW models, and are either 1 or 2 for the LP model. Notice that these results are consistent with our findings in Section 3.3.1: the propagation mechanism in the LP model is not strong enough to transmit the sector-specific shocks into common factors. Therefore, the common factor of X_t must originate from aggregate shocks, and this is why the estimated sectoral shocks $\hat{u}_t$ from this model contain common factors. On the other hand, although $\hat{u}_t$ estimated from the other two models are found to be sector specific, i.e., they do not share any common factors, the common factor of X_t can originate from some of the sector-specific shocks because the input-output linkages in these models are relatively strong.

TABLE 3.3: The Estimated Number of Static Factors for $\hat{u}_t$

Models	PC_1	PC_2	PC_3	IC_1	IC_2	IC_3	Onatski
Long and Plosser	2	2	8	1	1	4	2
Horvath	2	1	8	0	0	3	0
FSW	3	2	10	0	0	5	0

Next, it is interesting to know which sectors are more likely to be the key sectors — sectors whose productivity shocks can translate into common factors for X_t. Motivated by the simple example in Section 3.2.3, in Tables 3.4 and 3.5 we list the top 10 sectors ranked by their intensities as material providers measured by both $(\sum_{j=1}^{N} \gamma_{ij})$ and $(\sum_{j=1}^{N} \mathbf{1}(\gamma_{ij} > 0))$. Figures 3.5 and 3.6 plot the distributions of these 2 measures.

The first intensity measurement $\sum_{j=1}^{N} \gamma_{ij}$ reflects the total amount of materials provided by a sector to other sectors, and the second measurement $\sum_{j=1}^{N} \mathbf{1}(\gamma_{ij} > 0)$ shows how many sectors use materials from sector i. First, not surprising, these top key sectors include those whose relevance for the whole economy are commonly recognized, such as oil extraction, power generation, iron and steel productions, and semiconductors. Second, as the distributions of these measures show, there are only a few important sectors in terms of either of those two intensity measures.

TABLE 3.4: Top 10 Sectors Ranked by $\sum_{j=1}^{N} \gamma_{ij}$

Sector Names	Codes	$\sum_{j=1}^{N} \gamma_{ij}$
Iron and Steel Products	3311.2	3.31
Semiconductors and Other Electronic Components	3344	2.87
Plastics Products	3261	2.38
Organic Chemicals	32511.9	2.35
Electr Power Generation	2211	2.16
Oil and Gas Extraction	211	1.84
Paper and Paperboard Mills	32212.3	1.81
Sawmills and Wood Preservation	3211	1.66
Resins and Synthetic Rubber	32521	1.59
Motor Vehicle Parts	3363	1.56

TABLE 3.5: Top 10 Sectors Ranked by $\sum_{j=1}^{N} \mathbf{1}(\gamma_{ij} > 0)$

Sector Names	Codes	$\sum_{j=1}^{N} \mathbf{1}(\gamma_{ij} > 0)$
Electr Power Generation	2211	115
Plastics Products	3261	109
Natural Gas Distribution	2212	107
Machine Shops	3327	105
Coating, Engraving, Heat Treating, and Allied Activities	3328	102
Semiconductors and Other Electronic Components	3344	92
Electrical Equipment	3353	87
Other Chemical Product and Preparation	3259	85
Paperboard Containers	32221	80
Other Fabricated Metal Products	3329	69

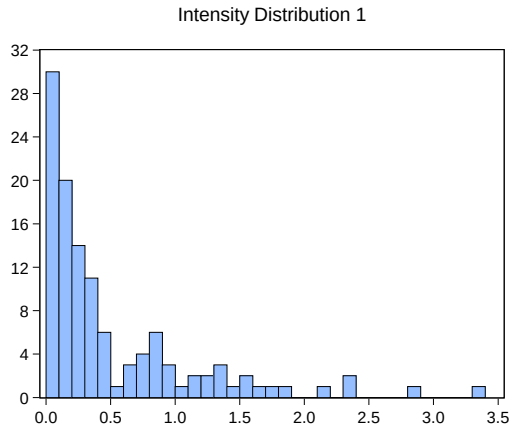


FIGURE 3.5: Distribution of $\sum_{j=1}^{N} \gamma_{ij}$

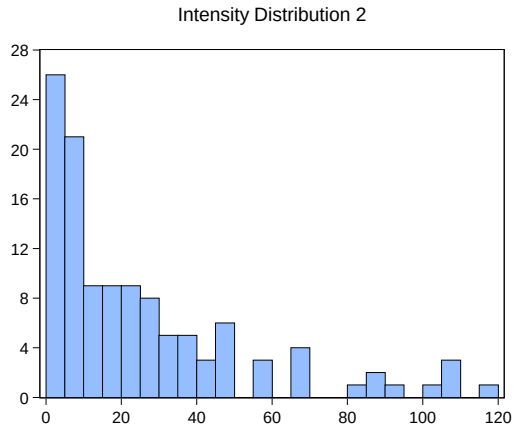


FIGURE 3.6: Distribution of $\sum_{j=1}^{N} \mathbf{1}(\gamma_{ij} > 0)$

As already mentioned, if the sectoral shocks satisfy Assumption 17, then we can use Proposition 1 to identify the key sectors and thus the common factors of X_t. More specifically, we can first divide the sectors into two groups — the key sectors and the other sectors, and partition the matrix $D(L)$ accordingly. Then we can calculate the eigenvalues of $\mathcal{A}(\omega)\tilde{\mathcal{A}}(\omega)$ and $\mathcal{B}(\omega)\tilde{\mathcal{B}}(\omega) - D_0\tilde{D}_0$. If we split the sectors correctly, the first eigenvalue of $\mathcal{A}(\omega)\tilde{\mathcal{A}}(\omega)$ will be much larger than the first eigenvalue of $\mathcal{B}(\omega)\tilde{\mathcal{B}}(\omega) - D_0\tilde{D}_0$. However, in practice, the sectoral shocks can be serially correlated, and can be weakly cross-sectionally correlated even though they do not share any common factors. Therefore the assumption of white noise could be possibly violated and this method does not work.

Given these shortcomings, to identify the common factors, we apply the method of Chen (2013), which is based on penalized regressions of the estimated common factors on different sets of observed variables. Suppose F_t is the common factors of X_t, and $F_t = \Upsilon z_t$, where z_t is a vector of m observed variables, and Υ is a $r \times m$ matrix. In other words, we assume that the common factors of X_t are linear combinations of some observed variables, which are denoted as *observed factors*. Moreover, suppose that z_t belongs to a set of M variables Z_t with $M >> m$ (without loss of generality, assume $z_t = [Z_{1t}, \ldots, Z_{mt}]'$). Then, the question is how to find out the observed factors z_t when we only observe X_t and Z_t. Chen (2013) has proposed the following two-step procedure: (i) in the first step, the common factors are estimated using PCA; (ii) in the second step, let $i_1 : i_k$ denote the set of k indices $[i_1, i_2, \ldots, i_k]$ with $1 \leq i_1 < i_2 \ldots < i_k \leq M$, and let $Z_{i_1:i_k,t}$ denote the vector $[Z_{i_1,t}, \ldots, Z_{i_k,t}]'$, then the set of indices are chosen to minimize the following object function:

$$S(i_1, i_k, k) = \frac{1}{T}\sum_{t=1}^{T}\left\|\hat{F}_t - \hat{\Upsilon}_k Z_{i_1:i_k,t}\right\|^2 + k \cdot p(N,T),$$

where $\hat{F}_t$ is the estimated factors, $\hat{\Upsilon}_k$ is the OLS estimates of the coefficients, and $p(N,T)$ is a penalty function that depends on N and T. If $p(N,T)$ satisfies $p(N,t) \to 0$ and $\min[N,T] \cdot p(N,T) \to \infty$ as $N,T \to \infty$, then we can consistently identify the observed factors in the following sense: $P[\hat{k} = m, \hat{i}_1 : \hat{i}_{\hat{k}} = 1 : m] \to 1$ as $N,T \to \infty$, where $\hat{k}$ and $\hat{i}_1 : \hat{i}_{\hat{k}}$ are the indices minimizing the above object function.

In our data set, X_t is shown to have only 1 common factor from 1984 to 2007. Further, in the structural models we consider, the only source of volatility are the sectoral shocks. Thus, the common factor of the sectoral IP growth rates can only originate from the latter. In the simple example where we assumed one key sector in the LP model, the common factor is connected with the shock to the key sector in a simple way: $F_t = X_{1,t-1} = (1-\gamma L)^{-1}u_{1t}$. However, in general, the common factor is expected to be associated with some of the sectoral shocks in a more complex way, so the candidates Z_t from which we search for variables whose linear combinations can approximate F_t should include $\hat{u}_t, \hat{u}_{t-1}, \ldots, \hat{u}_{t-p}$.

In Table 3.6 we report the identification results for the three models using the Chen' (2013) approach. The candidates Z_t include $\hat{u}_t$ and their lags up to order 4. We also consider the

TABLE 3.6: Identified Observed Factors Using Structural Models

	Long and Plosser 1983					
	$r_u = 0$			$r_u = 1$		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
Sector codes	81	81,116	64,81,116	g_t	g_t, g_{t-1}	g_t, g_{t-1}, g_{t-2}
p_1	0.5603	0.4742	0.4770	0.4657	**0.3279**	0.3384
p_2	0.5717	0.4969	0.5111	0.4770	**0.3507**	0.3725
p_3	0.5327	0.4189	0.3941	0.4380	0.2726	**0.2555**
	Horvath 1998					
	$r_u = 0$			$r_u = 1$		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
Sector codes	81	69,81	69,75,81	g_t	g_t, g_{t-1}	69,81,g_t
p_1	0.7614	0.6136	0.5821	0.5573	0.4786	**0.4695**
p_2	0.7728	0.6363	0.6153	0.5687	**0.5013**	0.5016
p_3	0.7338	0.5583	0.4982	0.5297	0.4233	**0.3846**
	FSW 2011					
	$r_u = 0$			$r_u = 1$		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
Sector codes	81	69,81	69,75,81	g_t	80,g_t	32,37,g_t
p_1	0.7120	0.6078	0.5404	**0.2737**	0.3057	0.3517
p_2	0.7234	0.6305	0.5744	**0.2850**	0.3284	0.3858
p_3	0.6843	0.5525	0.4574	**0.2460**	0.2504	0.2887

Codes and Sector Names: [32]Apparel; [37]Millwork; [64]Glass and Glass Products; [69] Iron and Steel Products; [75]Fabricated Metals: Cutlery and Hand tools; [80]Machine Shops; [81]Coating, Engraving, Heat Treating, and Allied Activities; [116]Newspaper Publishers.

case where u_t has a aggregate shock g_t: $u_t = Hg_t + \xi_t$. In this case, in addition to the sector specific shocks ξ_t and their lags, the common factor g_t is also included in the list of candidates. We use the following three versions of the penalty function:

$$p_1(N,T) = \left(\frac{N+T}{NT}\right)\ln\left(\frac{NT}{N+T}\right),\ p_2(N,T) = \left(\frac{N+T}{NT}\right)\ln\left(\delta^2_{N,T}\right),\ p_3(N,T) = \frac{\ln\delta^2_{N,T}}{\delta^2_{N,T}}.$$

They all satisfy the conditions stated above, but perform differently in finite samples (see Chen, 2013 for details). There are three panels in Table 3.6 corresponding to the three different models we consider. The first row of each panel reports the selected observed factors for $k = 1, 2, 3$. The second to last rows reported values of the object function for different choices of the penalty function, where the numbers in bold indicate the best choices of the observed factors, whose linear combinations can best approximate the common factor of X_t.

From Table 3.6 we can see that in all three models the common factor of X_t is better approximated by the aggregate shock of u_t: g_t. In particular, the LP and FSW models attribute the comovement of X_t exclusively to g_t while, in the H model, the common factor

F_t is also correlated with two sector-specific shocks (ξ_{it} from sectors 69 and 81)[4]. Moreover, when we ignore the penalty function and compare the results for each k (the penalty functions may not work in these cases due to the specification errors of the models), these two sector-specific shocks (69 and 81) are identified as proxies for the common factor F_t in most cases. Interestingly, these two sectors are indeed among the top 10 key sectors listed in Tables 3.4 and 3.5 in terms of intensities as input providers.

In sum, the empirical results strongly support the existence of an aggregate technology shock as the main source of the aggregate volatility, which could also be affected by the sector-specific shocks from two key sectors through input-output linkages.

3.4 Conclusion

In this paper we study how aggregate volatility can be determined by sector-specific shocks through the lens of dynamic factor models. Three structural models building on the assumption of multiple producing sectors are considered — Long and Plosser (1983), Horvath (1998), and Foerster et al (2011). We use a simple example to illustrate the point that: in a economy with only sector-specific shocks, the aggregate volatility is mainly determined by the shocks of the key sectors whose outputs are used by many other sectors as inputs.

Using data on the input-output matrix and the capital-use matrix, we show that even mutually independent sectoral shocks can be propagated into common factors through the input-output linkages in the last 2 models. While in the first model, such linkages are not strong enough so the effects of the sector-specific shocks will be averaged out when calculating the aggregate volatility. Using data on the industrial productions for 117 4-digit sectors in the US, we find that after the great moderation (1983) the IP growth rates of these 117 sectors can be well described by an approximate factor model with only 1 common factor. To study how this common factor is connected to aggregate and sector specific shocks, we use a regression-based method to identify this common factor. We find that the common factor is primarily connected to a aggregate technology shock that affects most of the sectors, and possibly to 1 or 2 sector-specific shocks that only affect the key sectors.

[4]For all of the 3 models, in the regressions of $\hat{F}_t$ on the selected variables, the R^2 from 80% to 90%

Appendix A

Appendix to Chapter 1

A.1 Proof of Propositions 1.1 and 1.2

The proof proceeds by showing that the errors, factors and loadings in model (1.5) satisfy Assumptions A to D of BN (2002). Then, once these results are proven, Propositions 1.1 and 1.2 just follow immediately from application of Theorems 1 and 2 of BN (2002). Define $F_t^* = [F_t' \quad G_t^{1'}]'$, $\epsilon_t = HG_t^2 + e_t$, and $\Gamma = [A \quad \Lambda]$. The proofs of Lemma A.1 to Lemma A.8 are similar to those in BN (2002). To save space and avoid repetition, we put them in our online appendix (http://www.eco.uc3m.es/ jgonzalo/WP1.html).

Lemma A.1. *$E||F_t^*||^4 < \infty$ and $T^{-1}\sum_{t=1}^T F_t^* F_t^{*'} \xrightarrow{p} \Sigma_F^*$ as $T \to \infty$ for some positive matrix Σ_F^*.*

Lemma A.2. *$||\Gamma_i|| < \infty$ for all i, and $N^{-1}\Gamma'\Gamma \to \Sigma_\Gamma$ as $N \to \infty$ for some positive definite matrix Σ_Γ.*

The following lemmae involve the new errors ϵ_t. Let M and M^* denote some positive constants.

Lemma A.3. *$E(\epsilon_{it}) = 0$, $E|\epsilon_{it}|^8 \leq M^*$.*

Lemma A.4. *$E(\epsilon_s'\epsilon_t/N) = E(N^{-1}\sum_{i=1}^N \epsilon_{is}\epsilon_{it}) = \gamma_N^*(s,t)$, $|\gamma_N^*(s,s)| \leq M^*$ for all s, and $\sum_{s=1}^T \gamma_N^*(s,t)^2 \leq M^*$ for all t and T.*

Lemma A.5. *$E(\epsilon_{it}\epsilon_{jt}) = \tau_{ij,t}^*$ with $|\tau_{ij,t}^*| \leq |\tau_{ij}^*|$ for some τ_{ij}^* and for all t; and $N^{-1}\sum_{i=1}^N \sum_{j=1}^N |\tau_{ij}^*| \leq M^*$.*

Lemma A.6. *$E(\epsilon_{it}\epsilon_{js}) = \tau_{ij,ts}^*$ and $(NT)^{-1}\sum_{i=1}^N \sum_{j=1}^N \sum_{t=1}^T \sum_{s=1}^T |\tau_{ij,ts}^*| \leq M^*$.*

Lemma A.7. *For every (t,s), $E|N^{-1/2}\sum_{i=1}^N [\epsilon_{is}\epsilon_{it} - E(\epsilon_{is}\epsilon_{it})]|^4 \leq M^*$.*

Lemma A.8. *$E\Big(\frac{1}{N}\sum_{i=1}^N \Big\|\frac{1}{\sqrt{T}}\sum_{t=1}^T F_t^*\epsilon_{it}\Big\|^2\Big) \leq M^*$.*

Finally, it is easy to verify that $\sum_{i=1}^{N}\sum_{t=1}^{T}\|\eta_i\|^2 E\|F_t\|^2 = O(1)$ and thus the new idiosyncratic errors ϵ_t satisfy the necessary condition for the consistency of $\hat{r}$ (See Observation 1 of Bates et al. 2013).

Once it has been shown that the new factors: F_t^*, the new loadings: Γ and the new errors: ϵ_t all satisfy the conditions of BN (2002), Propositions 1.1 and 1.2 just follow directly from their Theorems 1 and 2, with r replaced by $r + k_1$ and F_t replaced by F_t^*.

A.2 Proof of Theorem 1.3

Under the null: $k_1 = 0$, when $\bar{r} = r$ we have

$$\hat{F}_t = DF_t + o_p(1).$$

Let $D_{(i\cdot)}$ denote the ith row of D, and $D_{(\cdot j)}$ denote the jth column of D. Define $\hat{\mathcal{F}}_t = DF_t$, and $\hat{\mathcal{F}}_{kt} = D_{(k\cdot)} \times F_t$ as the kth element of $\hat{\mathcal{F}}_t$. Let $\hat{F}_{1t}$ be the first element of $\hat{F}_t$, and $\hat{F}_{-1t} = [\hat{F}_{2t}, \ldots, \hat{F}_{rt}]'$, while $\hat{\mathcal{F}}_{1t}$ and $\hat{\mathcal{F}}_{-1t}$ can be defined in the same way. Note that $\hat{\mathcal{F}}_t$ depends on N and T. For simplicity, let $T\pi$ denote $[T\pi]$.

Note that under H_0, we allow for the existence of small breaks, so that the model can be written as $X_{it} = \alpha_i F_t + e_{it} + \eta_i G_t^2$. However, since $\eta_i G_t^2$ is $O_p(1/\sqrt{NT})$ by Assumption 1, we can use similar methods as in Appendix A.1 to show that an error term of this order can be ignored and that the asymptotic properties of $\hat{F}_t$ will not be affected (See Remark 5 of Bai, 2009). Therefore, for the sake of simplicity in the presentation below, we eliminate the last term and consider instead the model $X_{it} = \alpha_i F_t + e_{it}$ in the following lemmae (A.9 to A.13) required to prove Lemma A.14, which is the key result in the proof of Theorem 1.3.

Lemma A.9.

$$\sup_{\pi\in[0,1]} \left\| \frac{1}{T}\sum_{t=1}^{T\pi} (\hat{F}_t - \hat{\mathcal{F}}_t) F_t' \right\| = O_p(\delta_{N,T}^{-2}).$$

Proof. The proof is similar to Lemma B.2 of Bai (2003). For details see our online appendix. □

Lemma A.10.

$$\sup_{\pi\in[0,1]} \left\| \frac{1}{T}\sum_{t=1}^{T\pi} \hat{F}_t \hat{F}_t' - \frac{1}{T}\sum_{t=1}^{T\pi} \hat{\mathcal{F}}_t \hat{\mathcal{F}}_t' \right\| = O_p(\delta_{N,T}^{-2}).$$

Proof. Note that:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T\pi}\hat{F}_t\hat{F}_t' - \frac{1}{T}\sum_{t=1}^{T\pi}\hat{\mathcal{F}}_t\hat{\mathcal{F}}_t' \\
= \;& \frac{1}{T}\sum_{t=1}^{T\pi}\hat{F}_t\hat{F}_t' - \frac{1}{T}\sum_{t=1}^{T\pi}(DF_t)(F_t'D') \\
= \;& \frac{1}{T}\sum_{t=1}^{T\pi}\hat{F}_t(\hat{F}_t' - F_t'D') + \frac{1}{T}\sum_{t=1}^{T\pi}(\hat{F}_t - DF_t)(F_t'D') \\
= \;& \frac{1}{T}\sum_{t=1}^{T\pi}(\hat{F}_t - DF_t)(\hat{F}_t - DF_t)' + \frac{1}{T}D\sum_{t=1}^{T\pi}F_t(\hat{F}_t - DF_t)' + \frac{1}{T}\sum_{t=1}^{T\pi}(\hat{F}_t - DF_t)(F_t'D').
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sup_{\pi\in[0,1]}\left\|\frac{1}{T}\sum_{t=1}^{T\pi}\hat{F}_t\hat{F}_t' - \frac{1}{T}\sum_{t=1}^{T\pi}\hat{\mathcal{F}}_t\hat{\mathcal{F}}_t'\right\| \\
\leq \;& \sup_{\pi\in[0,1]}\left\|\frac{1}{T}\sum_{t=1}^{T\pi}(\hat{F}_t - DF_t)(\hat{F}_t - DF_t)'\right\| + 2\|D\|\sup_{\pi\in[0,1]}\left\|\frac{1}{T}\sum_{t=1}^{T\pi}(\hat{F}_t - DF_t)F_t'\right\| \\
\leq \;& \frac{1}{T}\sum_{t=1}^{T}\left\|\hat{F}_t - DF_t\right\|^2 + 2\|D\|\sup_{\pi\in[0,1]}\left\|\frac{1}{T}\sum_{t=1}^{T\pi}(\hat{F}_t - DF_t)F_t'\right\|.
\end{aligned}$$

Since $\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{F}_t - DF_t\right\|^2 = O_p(\delta_{N,T}^{-2})$ and $\sup_{\pi\in[0,1]}\left\|\frac{1}{T}\sum_{t=1}^{T\pi}(\hat{F}_t - DF_t)F_t'\right\|$ is $O_p(\delta_{N,T}^{-2})$ by Lemma A.9, then the desired result is obtained.

□

The next two lemmae follow from Lemma A.10 and Assumption 6:

Lemma A.11.

$$\sup_{\pi\in[0,1]}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\hat{F}_{-1t}\hat{F}_{1t} - \frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\hat{\mathcal{F}}_{-1t}\hat{\mathcal{F}}_{1t}\right\| = o_p(1).$$

Proof. See Lemma A.10 and Assumption 6. □

Lemma A.12.

$$\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\hat{\mathcal{F}}_{-1t}\hat{\mathcal{F}}_{1t}'\right\| = o_p(1).$$

Proof. By construction we have $\frac{1}{T}\sum_{t=1}^{T}\hat{F}_{-1t}\hat{F}_{1t}' = 0$, and then the result follows from Lemma A.11. □

Let $\Rightarrow$ denote *weak convergence*. $\mathcal{F}_{1t}$, $\mathcal{F}_{-1t}$, D^* and S are as defined in the paper (Page 12). Then:

Lemma A.13.

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\left(\mathcal{F}_{-1t}\mathcal{F}_{1t} - E(\mathcal{F}_{-1t}\mathcal{F}_{1t})\right) \Rightarrow S^{1/2}\mathcal{W}_{r-1}(\pi)$$

for $\pi \in [0,1]$, *where* $\mathcal{W}_{r-1}(\cdot)$ *is a* $r-1$ *vector of independent Brownian motions on* $[0,1]$.

Proof. The proof is a standard application of Functional CLT. For details see our online appendix. □

Lemma A.14.

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\hat{F}_{-1t}\hat{F}_{1t} \Rightarrow S^{1/2}\mathcal{B}^0_{r-1}(\pi)$$

for $\pi \in [0,1]$*, where the process* $\mathcal{B}^0_{r-1}(\pi) = \mathcal{W}_{r-1}(\pi) - \pi\mathcal{W}_{r-1}(1)$ *indexed by* π *is a vector of Brownian Bridges on* $[0,1]$*.*

Proof. If we show that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\left[\mathcal{F}_{-1t}\mathcal{F}_{1t} - T^{-1}\sum_{s=1}^{T}\mathcal{F}_{-1s}\mathcal{F}_{1s}\right] \Rightarrow S^{1/2}\mathcal{B}^0_{r-1}(\pi) \tag{A.1}$$

for $\pi \in [0,1]$ and

$$\sup_{\pi\in[0,1]}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\hat{\mathcal{F}}_{-1t}\hat{\mathcal{F}}_{1t} - \frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\left[\mathcal{F}_{-1t}\mathcal{F}_{1t} - T^{-1}\sum_{s=1}^{T}\mathcal{F}_{-1s}\mathcal{F}_{1s}\right]\right\| = o_p(1), \tag{A.2}$$

then the result follows from Lemma A.11.

First note that

$$\begin{aligned}
&\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\left[\mathcal{F}_{-1t}\mathcal{F}_{1t} - T^{-1}\sum_{s=1}^{T}\mathcal{F}_{-1s}\mathcal{F}_{1s}\right] \\
= \;&\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\left(\mathcal{F}_{-1t}\mathcal{F}_{1t} - E(\mathcal{F}_{-1t}\mathcal{F}_{1t})\right) + \frac{1}{T}\sum_{t=1}^{T\pi}\left(\frac{1}{\sqrt{T}}\sum_{s=1}^{T}\left(\mathcal{F}_{-1s}\mathcal{F}_{1s} - E(\mathcal{F}_{-1s}\mathcal{F}_{1s})\right)\right),
\end{aligned}$$

hence (A.1) can be verified by applying Lemma A.13.

To prove (A.2), we first define D_{-1} as the second to last rows of D, and D_1 as the first row of D. D^*_{-1} and D^*_1 are defined in the same manner. Then we have

$$\hat{\mathcal{F}}_{-1t}\hat{\mathcal{F}}_{1t} = D_{-1}F_tF_t'D_1'$$

and

$$\mathcal{F}_{-1t}\mathcal{F}_{1t} = D^*_{-1}F_tF_t'D_1^{*'}.$$

It follows that:

$$\begin{aligned}
&\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\left(\hat{\mathcal{F}}_{-1t}\hat{\mathcal{F}}_{1t} - \mathcal{F}_{-1t}\mathcal{F}_{1t}\right) \\
= \;&\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\left(D_{-1}F_tF_t'D_1' - D_{-1}F_tF_t'D_1^* + D_{-1}F_tF_t'D_1^* - D^*_{-1}F_tF_t'D_1^*\right) \\
= \;&D_{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}F_tF_t'\right)\left(D_1' - D_1^{*'}\right) + \left(D_{-1} - D^*_{-1}\right)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}F_tF_t'\right)D_1^{*'}.
\end{aligned}$$

Next, define $\overline{FF'} = \frac{1}{T}\sum_{s=1}^{T} F_s F_s'$, then:

$$\begin{aligned}
&\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\Big(T^{-1}\sum_{s=1}^{s}\mathcal{F}_{-1s}\mathcal{F}_{1s}\Big)\\
=\;& D_{-1}^{*}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\overline{FF'}\right)D_1^{*'}\\
=\;& D_{-1}^{*}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\overline{FF'}\right)D_1^{*'} - D_{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\overline{FF'}\right)D_1^{*'} + D_{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\overline{FF'}\right)D_1^{*'}\\
&-D_{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\overline{FF'}\right)D_1' + D_{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\overline{FF'}\right)D_1'\\
=\;& -(D_{-1}-D_{-1}^{*})\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\overline{FF'}\right)D_1^{*'} - D_{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\overline{FF'}\right)(D_1'-D_1^{*'}) + \frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\Big(\frac{1}{T}\sum_{s=1}^{T}\hat{\mathcal{F}}_{-1s}\hat{\mathcal{F}}_{1s}\Big).
\end{aligned}$$

Combining the above results gives:

$$\begin{aligned}
&\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\hat{\mathcal{F}}_{-1t}\hat{\mathcal{F}}_{1t} - \frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\Big[\mathcal{F}_{-1t}\mathcal{F}_{1t} - T^{-1}\sum_{s=1}^{T}\mathcal{F}_{-1s}\mathcal{F}_{1s}\Big]\\
=\;& \frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\left(\hat{\mathcal{F}}_{-1t}\hat{\mathcal{F}}_{1t} - \mathcal{F}_{-1t}\mathcal{F}_{1t}\right) + \frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\Big(T^{-1}\sum_{s=1}^{T}\mathcal{F}_{-1s}\mathcal{F}_{1s}\Big)\\
=\;& D_{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\left(F_tF_t' - \overline{FF'}\right)\right)(D_1'-D_1^{*'}) + (D_{-1}-D_{-1}^{*})\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\left(F_tF_t' - \overline{FF'}\right)\right)D_1^{*'}\\
&+\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\Big(\frac{1}{T}\sum_{s=1}^{T}\hat{\mathcal{F}}_{-1s}\hat{\mathcal{F}}_{1s}\Big).
\end{aligned}$$

Following similar arguments as in Lemma A.13, it can proved that

$$\sup_{\pi\in[0,1]}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T\pi}\left(F_tF_t' - \overline{FF'}\right)\right\| = O_p(1).$$

Moreover, it is easy to check that $\|D\| = O_p(1)$ and $\|D - D^*\| = o_p(1)$ (See Bai 2003). Finally, $\left\|\frac{1}{\sqrt{T}}\sum_{s=1}^{T}\hat{\mathcal{F}}_{-1s}\hat{\mathcal{F}}_{1s}\right\|$ is $o_p(1)$ by Lemma A.12. Then (A.2) holds and we obtain the desired conclusion. □

Theorem 1.3:

Proof. The results for LM and Sup-LM tests follow from Assumption 9, Lemma A.14, and Continuous Mapping Theorem.

For the Wald and Sup-Wald tests, notice that:

$$\sqrt{T}\big(\hat{c}_1(\pi)-\hat{c}_2(\pi)\big) = \big(1/T\sum_{t=1}^{\tau}\hat{F}_{-1t}\hat{F}_{-1t}'\big)^{-1}\big(1/\sqrt{T}\sum_{t=1}^{\tau}\hat{F}_{-1t}\hat{F}_{1t}\big) - \big(1/T\sum_{t=\tau+1}^{T}\hat{F}_{-1t}\hat{F}_{-1t}'\big)^{-1}\big(1/\sqrt{T}\sum_{t=\tau+1}^{T}\hat{F}_{-1t}\hat{F}_{1t}\big)$$

$$= \left[\big(1/T\sum_{t=1}^{\tau}\hat{F}_{-1t}\hat{F}_{-1t}'\big)^{-1} + \big(I - 1/T\sum_{t=1}^{\tau}\hat{F}_{-1t}\hat{F}_{-1t}'\big)^{-1}\right]\big(1/\sqrt{T}\sum_{t=1}^{\tau}\hat{F}_{-1t}\hat{F}_{1t}\big).$$

By Lemma A.10 and that $D - D^* = o_p(1)$, we have:

$$1/T\sum_{t=1}^{\tau}\hat{F}_{-1t}\hat{F}'_{-1t} = \pi\frac{1}{\tau}\sum_{t=1}^{\tau}\hat{F}_{-1t}\hat{F}'_{-1t} = \pi\frac{1}{\tau}\sum_{t=1}^{\tau}\mathcal{F}_{-1t}\mathcal{F}'_{-1t} + o_p(1). \tag{A.3}$$

When $\tau = T$ ($\pi = \tau/T = 1$), this implies

$$I_{r-1} = 1/T\sum_{t=1}^{T}\hat{F}_{-1t}\hat{F}'_{-1t} = \frac{1}{T}\sum_{t=1}^{T}\mathcal{F}_{-1t}\mathcal{F}'_{-1t} + o_p(1).$$

Notice that $E(\mathcal{F}_{-1t}\mathcal{F}'_{-1t}) = I_{r-1}$, because $E(\mathcal{F}_t\mathcal{F}'_t) = D^*\Sigma_F D^* = V^{-1/2}\Gamma'\Sigma_\Lambda^{1/2}\Sigma_F\Sigma_\Lambda^{1/2}\Gamma V^{-1/2} = I_r$. Applying law of large numbers to (A.3) gives:

$$1/T\sum_{t=1}^{\tau}\hat{F}_{-1t}\hat{F}'_{-1t} \xrightarrow{p} \pi I_{r-1}$$

as N and T go to infinity. Then it follows from Lemma A.14 that:

$$\sqrt{T}\big(\hat{c}_1(\pi) - \hat{c}_2(\pi)\big) \Rightarrow \frac{S^{1/2}\mathcal{B}^0_{r-1}(\pi)}{\pi(1-\pi)}$$

and the limit distributions of the Wald and Sup-Wald tests follow easily. □

A.3 Consistent Estimator of S

We now discuss the consistent estimator of S using the HAC estimator of Newey and West (1987). Recall that $S = \lim \mathrm{Var}\Big(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathcal{F}_{-1t}\mathcal{F}_{1t}\Big)$. Notice that $E(\mathcal{F}_{-1t}\mathcal{F}_{1t}) = 0$ as shown above.

First, we define the infeasible estimator of S:

$$\hat{S}(\mathcal{F}) = \hat{\Gamma}_0(\mathcal{F}) + \sum_{j=1}^{m} w(j,m)[\hat{\Gamma}_j(\mathcal{F}) + \hat{\Gamma}_j(\mathcal{F})']$$

where $m = o(T^{\frac{1}{4}})$, $w(j,m) = 1 - \frac{j}{m+1}$ is the Bartlett kernel, and

$$\hat{\Gamma}_j(\mathcal{F}) = \frac{1}{T}\sum_{t=j+1}^{T}\mathcal{F}_{-1t}\mathcal{F}_{1t}\mathcal{F}_{1t-j}\mathcal{F}'_{-1t-j}.$$

Since the above estimator is a HAC estimator, it is natural to make the following assumption:

Assumption 19. $\big\|\hat{S}(\mathcal{F}) - S\big\| = o_p(1)$.

Next we consider a feasible estimator of S where $\mathcal{F}_t$ is replaced by $\hat{F}_t$:

$$\hat{S}(\hat{F}) = \hat{\Gamma}_0(\hat{F}) + \sum_{j=1}^{m} w(j,m)[\hat{\Gamma}_j(\hat{F}) + \hat{\Gamma}_j(\hat{F})']$$

where

$$\hat{\Gamma}_j(\hat{F}) = \frac{1}{T}\sum_{t=j+1}^{T} \hat{F}_{-1t}\hat{F}_{1t}\hat{F}_{1t-j}\hat{F}'_{-1t-j}.$$

then we have the following results:

Proposition A.15. *Assume that Assumptions 1 to 10 and 19 hold, under the null $H_0 : k_1 = 0$, we have*

$$\left\|\hat{S}(\hat{F}) - S\right\| = o_p(1).$$

Proof. Given Assumption 19, it suffices to show that

$$\left\|\hat{S}(\hat{F}) - \hat{S}(\mathcal{F})\right\| = o_p(1).$$

It is easy to see that:

$$\left\|\hat{S}(\hat{F}) - \hat{S}(\mathcal{F})\right\| \leq 2\sum_{j=0}^{m}\left\|\hat{\Gamma}_j(\hat{F}) - \hat{\Gamma}_j(\mathcal{F})\right\|.$$

For the right-hand side we obtain that:

$$\begin{aligned}
&\sup_{0\leq j\leq m}\left\|\hat{\Gamma}_j(\hat{F}) - \hat{\Gamma}_j(\mathcal{F})\right\| \\
\leq\ &\sup_{0\leq j\leq m}\frac{1}{T}\sum_{t=j+1}^{T}\left\|\hat{F}_{-1t}\hat{F}_{1t}\hat{F}_{1t-j}\hat{F}'_{-1t-j} - \mathcal{F}_{-1t}\mathcal{F}_{1t}\mathcal{F}_{1t-j}\mathcal{F}'_{-1t-j}\right\|
\end{aligned}$$

$$\leq \sup_{0\leq j\leq m}\frac{1}{T}\sum_{t=j+1}^{T}\left\|(\hat{F}_{1t}\hat{F}_{-1t} - \mathcal{F}_{1t}\mathcal{F}_{-1t})\hat{F}_{1t-j}\hat{F}'_{-1t-j}\right\| + \sup_{0\leq j\leq m}\frac{1}{T}\sum_{t=j+1}^{T}\left\|\mathcal{F}_{1t}\mathcal{F}_{-1t}(\hat{F}_{1t-j}\hat{F}'_{-1t-j} - \mathcal{F}_{1t-j}\mathcal{F}'_{-1t-j})\right\|$$

$$\leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{F}_{1t}\hat{F}_{-1t} - \mathcal{F}_{1t}\mathcal{F}_{-1t}\right\|^2}\left(\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{F}_{1t}\hat{F}_{-1t}\right\|^2} + \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left\|\mathcal{F}_{1t}\mathcal{F}_{-1t}\right\|^2}\right).$$

$\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{F}_{1t}\hat{F}_{-1t}\right\|^2 \leq \frac{1}{T}\sum_{t=1}^{T}\left\|\hat{F}_t\right\|^4 = O_p(1)$ by Lemma 5 of HI (2012), and $\frac{1}{T}\sum_{t=1}^{T}\left\|\mathcal{F}_{1t}\mathcal{F}_{-1t}\right\|^2 \leq \frac{1}{T}\sum_{t=1}^{T}\left\|\mathcal{F}_t\right\|^4 = O_p(1)$ by Assumption 2. Furthermore,

$$\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{F}_{1t}\hat{F}_{-1t} - \mathcal{F}_{1t}\mathcal{F}_{-1t}\right\|^2$$

$$\leq \frac{1}{T}\sum_{t=1}^{T}\left\|(\hat{F}_{1t} - \mathcal{F}_{1t})\hat{F}_{-1t}\right\|^2 + \frac{1}{T}\sum_{t=1}^{T}\left\|(\hat{F}_{-1t} - \mathcal{F}_{-1t})\mathcal{F}_{1t}\right\|^2$$

$$\leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{F}_t - \mathcal{F}_t\right\|^4}\left(\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{F}_t\right\|^4} + \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left\|\mathcal{F}_t\right\|^4}\right).$$

It can be proved that $\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{F}_t - \mathcal{F}_t\right\|^4 = O_p(1/T) + O_p(1/N^2)$ (very similar to Theorem 2 of HI, 2012). Then, under the assumption that $\sqrt{T}/N \to 0$, it follows that

$$\sup_{0\leq j\leq m}\frac{1}{T}\sum_{t=j+1}^{T}\left\|\hat{F}_{1t}\hat{F}_{-1t}\hat{F}_{1t-j}\hat{F}'_{-1t-j} - \mathcal{F}_{1t}\mathcal{F}_{-1t}\mathcal{F}_{1t-j}\mathcal{F}'_{-1t-j}\right\| = O_p(T^{-1/4}),$$

which implies $\left\|\hat{S}(\hat{F}) - \hat{S}(\mathcal{F})\right\| = o_p(1)$ given that $m = o(T^{1/4})$. □

Appendix B

Appendix B

Appendix to Chapter 2

B.1 Proof of Theorem 2.2

Lemma B.1. *(Bai 2003) Let* $\tilde{f} = (\tilde{f}_1, \ldots, \tilde{f}_T)'$, *and* $f = (f_1, \ldots, f_T)'$, *then under Assumptions 11 to 13,*

$$\tilde{f}' f / T \xrightarrow{p} Q,$$

where $Q = V^{1/2}\Gamma'\Sigma_\Lambda^{-1/2}$, V *is a diagonal matrix consisting of the eigenvalues of* $\Sigma_\Lambda^{1/2}\Sigma_F\Sigma_\Lambda^{1/2}$ *in decreasing order and* Γ *consists of the corresponding eigenvectors.*

Lemma B.2. *When* Q *is defined as in Lemma 2.5, then* $Q'Q = \Sigma_F$.

Proof. By definition, $Q'Q = \Sigma_\Lambda^{-1/2}\Gamma V^{1/2}V^{1/2}\Gamma'\Sigma_\Lambda^{-1/2} = \Sigma_\Lambda^{-1/2}\Gamma V\Gamma'\Sigma_\Lambda^{-1/2}$. Also we have $\Sigma_\Lambda^{1/2}\Sigma_F\Sigma_\Lambda^{1/2} = \Gamma V\Gamma'$, so $Q'Q = \Sigma_\Lambda^{-1/2}\big(\Sigma_\Lambda^{1/2}\Sigma_F\Sigma_\Lambda^{1/2}\big)\Sigma_\Lambda^{-1/2} = \Sigma_F$. □

Now let's consider a set of indices $n_1 : n_r = (n_1, \ldots, n_r)$, and the corresponding observed variables $X_{n_1:n_r,t} = (x_{n_1,t}, \ldots, x_{n_r,t})'$. We can write:

$$X_{n_1:n_r,t} = \Lambda_{n_1:n_r} f_t + e_{n_1:n_r,t}.$$

We have seen that $\min_A S(n_1 : n_r, A) = S(n_1 : n_r, \hat{A})$, where $\hat{A}' = [\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_r]$, and $\hat{A}_k$ is the OLS estimator of A_k. For simplicity, we use $S(n_1 : n_r)$ to denote $S(n_1 : n_r, \hat{A})$ in the sequel which is equal to:

$$\frac{1}{T}\mathrm{Tr}\Big[\tilde{f}'\big(I_T - X_{n_1:n_r}(X'_{n_1:n_r}X_{n_1:n_r})^{-1}X'_{n_1:n_r}\big)\tilde{f}\Big], \tag{B.1}$$

where $X_{n_1:n_r} = (X_{n_1:n_r,1}, \ldots, \mathbf{x}_{n_1:n_r,T})'$. The following result is key to prove Theorem 2.2.

Lemma B.3. *Under Assumptions 11 to 14:*

$$S(n_1 : n_r) \xrightarrow{p} Tr\Big[\big(\Lambda_{n_1:n_r}\Sigma_F\Lambda'_{n_1:n_r} + \Sigma^e_{n_1:n_r}\big)^{-1}\Sigma^e_{n_1:n_r}\Big] \tag{B.2}$$

where $\Sigma^e_{n_1:n_r} = plim\frac{1}{T}\sum_{t=1}^T e_{n_1:n_r,t}e'_{n_1:n_r,t}$.

Proof. We have

$$\begin{aligned}
&\frac{1}{T}\tilde{f}'\big(I_T - X_{n_1:n_r}(X'_{n_1:n_r}X_{n_1:n_r})^{-1}X'_{n_1:n_r}\big)\tilde{f} \\
= \;& \frac{1}{T}\tilde{f}'\tilde{f} - \Big(\frac{\tilde{f}'X_{n_1:n_r}}{T}\Big)\Big(\frac{X'_{n_1:n_r}X_{n_1:n_r}}{T}\Big)^{-1}\Big(\frac{X'_{n_1:n_r}\tilde{f}}{T}\Big) \\
= \;& I_r - \Big(\frac{\tilde{f}'X_{n_1:n_r}}{T}\Big)\Big(\frac{X'_{n_1:n_r}X_{n_1:n_r}}{T}\Big)^{-1}\Big(\frac{X'_{n_1:n_r}\tilde{f}}{T}\Big).
\end{aligned}$$

One can write $X_{n_1:n_r} = f\Lambda'_{n_1:n_r} + e_{n_1:n_r}$, where $e_{n_1:n_r} = (e_{n_1:n_r,1}, \ldots, e_{n_1:n_r,T})'$. Then:

$$\begin{aligned}
\frac{\tilde{f}'X_{n_1:n_r}}{T} &= \frac{\tilde{f}'f}{T}\Lambda'_{n_1:n_r} + \frac{\tilde{f}'e_{n_1:n_r}}{T} \\
&= \frac{\tilde{f}'f}{T}\Lambda'_{n_1:n_r} + H'\frac{f'e_{n_1:n_r}}{T} + \frac{\big(\tilde{f} - fH\big)'e_{n_1:n_r}}{T}.
\end{aligned}$$

Firstly, $\frac{\tilde{f}'f}{T}$ converges in probability to Q by Lemma 2.5. Secondly, $\frac{f'e_{n_1:n_r}}{T} = \frac{1}{T}\sum_{t=1}^{T} f_t e'_{n_1:n_r,t}$. If $1 \le i \le r$, then $\frac{1}{T}\sum_{t=1}^{T} f_{kt}e_{it} = o_p(1)$ by Assumption 14; if $i \ge r$, then $E|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f_{kt}e_{it}|^2 = \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} E(f_{ks}f_{kt})E(e_{it}e_{is}) \le \frac{C}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\gamma_{is} \le CM$ by Assumptions 11, 12 and 13, where C is a finite constant, and thus $\frac{1}{T}\sum_{t=1}^{T} f_{kt}e_{it}$ is $o_p(1)$. Moreover, we have $\|H\| = O_p(1)$ (See Bai 2003). Therefore $H'\frac{f'e_{n_1:n_r}}{T} = o_p(1)$. Finally, the last term is $o_p(1)$ by Lemma 2.1 and thus we have:

$$\frac{\tilde{f}'X_{n_1:n_r}}{T} \xrightarrow{p} Q\Lambda'_{n_1:n_r}. \tag{B.3}$$

Using similar arguments, we can show that:

$$\frac{X'_{n_1:n_r}X_{n_1:n_r}}{T} \xrightarrow{p} \Sigma^X_{n_1:n_r} = \Lambda_{n_1:n_r}\Sigma_F\Lambda'_{n_1:n_r} + \Sigma^e_{n_1:n_r}, \tag{B.4}$$

and it is easy to show that $\Sigma^X_{n_1:n_r} > 0$ and thus is invertible by Assumptions 11(i), 14(i) and 14(iii). Combining the above results we have:

$$\begin{aligned}
&\frac{1}{T}\tilde{f}'\big(I_T - X_{n_1:n_r}(X'_{n_1:n_r}X_{n_1:n_r})^{-1}X'_{n_1:n_r}\big)\tilde{f} \\
\xrightarrow{p} \;& I_r - Q\Lambda'_{n_1:n_r}\big(\Lambda_{n_1:n_r}\Sigma_F\Lambda'_{n_1:n_r} + \Sigma^e_{n_1:n_r}\big)^{-1}\Lambda_{n_1:n_r}Q',
\end{aligned}$$

and

$$\begin{aligned}
& S(n_1 : n_r) \\
\xrightarrow{p} \;& \mathrm{Tr}\Big[I_r - Q\Lambda'_{n_1:n_r}\big(\Lambda_{n_1:n_r}\Sigma_F\Lambda'_{n_1:n_r} + \Sigma^e_{n_1:n_r}\big)^{-1}\Lambda_{n_1:n_r}Q'\Big] \\
= \;& \mathrm{Tr}\Big[I_r - \big(\Lambda_{n_1:n_r}\Sigma_F\Lambda'_{n_1:n_r} + \Sigma^e_{n_1:n_r}\big)^{-1}\Lambda_{n_1:n_r}Q'Q\Lambda'_{n_1:n_r}\Big] \\
= \;& \mathrm{Tr}\Big[I_r - \big(\Lambda_{n_1:n_r}\Sigma_F\Lambda'_{n_1:n_r} + \Sigma^e_{n_1:n_r}\big)^{-1}\Lambda_{n_1:n_r}\Sigma_F\Lambda'_{n_1:n_r}\Big] \\
= \;& \mathrm{Tr}\Big[\big(\Lambda_{n_1:n_r}\Sigma_F\Lambda'_{n_1:n_r} + \Sigma^e_{n_1:n_r}\big)^{-1}\Sigma^e_{n_1:n_r}\Big],
\end{aligned}$$

as desired. □

To prove Theorem 2.2, notice that when the DOFs are selected, $n_1 : n_r = 1 : r = (1, 2, \ldots, r)$, and $\Sigma^e_{1:r} = 0$ by Assumption 14(i). Therefore, we have

$$\text{plim } S(1 : r) = 0.$$

While when we select the wrong set of variables, $\Sigma^e_{n_1:n_r}$ is either positive definite or positive semi-definite. It is positive semi-definite when part of the selected variables belong to the first r variables, i.e., when there exists at least one N_l such that $1 \leq N_l \leq r$, but it cannot be 0 as long as one of the selected variables does nott belong to DOFs. Then, by the fact that $\Lambda_{n_1:n_r}\Sigma_F\Lambda'_{n_1:n_r} + \Sigma^e_{n_1:n_r} > 0$, we have

$$\text{plim } S(n_1 : n_r) > 0.$$

Then Theorem 2.2 follows easily.

B.2 Proof of Theorem 2.3

Lemma B.4. $Q\Sigma_F^{-1}Q' = I_r$

Proof. By definition, $Q\Sigma_F^{-1}Q' = V^{1/2}\Gamma'\Sigma_\Lambda^{-1/2}\Sigma_F^{-1}\Sigma_\Lambda^{-1/2}\Gamma V^{1/2} = V^{1/2}\Gamma'\big(\Gamma V\Gamma'\big)^{-1}\Gamma V^{1/2} = I_r$ □

Lemma B.5. *For any $k > r$ and $m + 1 \leq n_1 < n_2 \ldots < N_k \leq N$, we have*

$$\textit{plim } S(n_1 : N_k) > 0.$$

Proof. By definition we can write:

$$X_{n_1:N_k,t} = \Lambda_{n_1:N_k} f_t + e_{n_1:N_k,t}.$$

Using the same arguments in Lemma B.2, we can show that:

$$S(n_1 : n_r) \xrightarrow{p} \text{Tr}\Big[I_r - Q\Lambda'_{n_1:N_k}\big(\Lambda_{n_1:N_k}\Sigma_F\Lambda'_{n_1:n_r} + \Sigma^e_{n_1:N_k}\big)^{-1}\Lambda_{n_1:N_k}Q'.\Big] \tag{B.5}$$

But now we can proceed as in the proof of Lemma B.2 because $\Lambda_{n_1:N_k}$ is not $r \times r$. Instead, by Lemma B.3, we can write the matrix in the right hand side of (B.5)as:

$$\begin{aligned} & Q\Sigma_F^{-1}Q' - Q\Lambda'_{n_1:N_k}\big(\Lambda_{n_1:N_k}\Sigma_F\Lambda'_{n_1:n_r} + \Sigma^e_{n_1:N_k}\big)^{-1}\Lambda_{n_1:N_k}Q' \\ = \; & Q\Big(\Sigma_F^{-1} - \Lambda'_{n_1:N_k}\big(\Lambda_{n_1:N_k}\Sigma_F\Lambda'_{n_1:n_r} + \Sigma^e_{n_1:N_k}\big)^{-1}\Lambda_{n_1:N_k}\Big)Q'. \end{aligned}$$

By Assumption 15(iii) and matrix inverse equation we have:

$$\begin{aligned} & \big(\Lambda_{n_1:N_k}\Sigma_F\Lambda'_{n_1:n_r} + \Sigma^e_{n_1:N_k}\big)^{-1} \\ = \; & (\Sigma^e_{n_1:N_k})^{-1} - (\Sigma^e_{n_1:N_k})^{-1}\Lambda_{n_1:n_r}\Big(\Sigma_F^{-1} + \Lambda'_{n_1:n_r}(\Sigma^e_{n_1:N_k})^{-1}\Lambda_{n_1:n_r}\Big)^{-1}\Lambda'_{n_1:n_r}(\Sigma^e_{n_1:N_k})^{-1}. \end{aligned}$$

Define $C = \Lambda'_{n_1:n_r}(\Sigma^e_{n_1:N_k})^{-1}\Lambda_{n_1:n_r}$, then we have:

$$\begin{aligned}
& \Sigma_F^{-1} - \Lambda'_{n_1:N_k}\big(\Lambda_{n_1:N_k}\Sigma_F\Lambda'_{n_1:n_r} + \Sigma^e_{n_1:N_k}\big)^{-1}\Lambda_{n_1:N_k} \\
= \; & \Sigma_F^{-1} - \Big(C - C\big(\Sigma_F^{-1} + C\big)^{-1}C\Big) \\
= \; & \Sigma_F^{-1} - \Big(C\big(\Sigma_F^{-1} + C\big)^{-1}\big(\Sigma_F^{-1} + C\big) - C\big(\Sigma_F^{-1} + C\big)^{-1}C\Big) \\
= \; & \Sigma_F^{-1} - C\big(\Sigma_F^{-1} + C\big)^{-1}\Sigma_F^{-1} \\
= \; & \big(\Sigma_F^{-1} + C\big)\big(\Sigma_F^{-1} + C\big)^{-1}\Sigma_F^{-1} - C\big(\Sigma_F^{-1} + C\big)^{-1}\Sigma_F^{-1} \\
= \; & \Sigma_F^{-1}\big(\Sigma_F^{-1} + C\big)^{-1}\Sigma_F^{-1}.
\end{aligned}$$

Finally we have:

$$\begin{aligned}
& S(n_1 : n_r) \\
\xrightarrow{p} \; & \mathrm{Tr}\Big[Q\Sigma_F^{-1}\big(\Sigma_F^{-1} + C\big)^{-1}\Sigma_F^{-1}Q'\Big] \\
= \; & \mathrm{Tr}\Big[\Sigma_F^{-1}\big(\Sigma_F^{-1} + C\big)^{-1}\Big]
\end{aligned}$$

by Lemma B.1. Then the result follows by the fact that both Σ_F and $\Sigma_F^{-1}+C$ are positive definite. □

Lemma B.6. *If $e'e$ is the sum of squared residuals when y is regressed on X and $u'u$ is the sum of squared residuals when y is regressed on X and z, then*

$$u'u = e'e - c^2(z'_*z_*) \leq e'e,$$

where c is the coefficient on z in the long regression and $z_* = [I - X(X'X)^{-1}X']z$ is the vector of residuals when z is regressed on X.

Proof. See Green (2002). □

Lemma B.5 implies that in OLS regressions, adding regressors never increases the RSS.

Lemma B.7. $S(1 : m) = O_p(\delta_{N,T}^{-2})$.

Proof. By Lemma 2.1 and Assumption 15(i), we have:

$$\tilde{f}_t = Hf_t + V_t = HBX_{1:m,t} + V_t = AX_{1:m,t} + V_t, \tag{B.6}$$

where $A = HB$ and $V_t = O_p(\delta_{N,T}^{-1})$. Then we can write:

$$\tilde{f}_t = \hat{A}X_{1:m,t} + (A - \hat{A})X_{1:m,t} + V_t.$$

Since

$$A - \hat{A} = \Big(T^{-1}\sum_{t=1}^{T} X_{1:m,t}X'_{1:m,t}\Big)^{-1}\Big(T^{-1}\sum_{t=1}^{T} X_{1:m,t}V'_t\Big) = O_p(\delta_{N,T}^{-1})$$

by Assumption 15(ii). It follows that

$$\|\tilde{f}_t - \hat{A}X_{1:m,t}\|^2 = O_p(\delta_{N,T}^{-2})$$

and the result follows. □

The following lemma states that if the IOFs are selected together with some other variables, the RSS divided by T also goes to 0.

Lemma B.8. *Let* $[1 : m, n_1 : N_l] = [1, 2 \ldots, m, n_1, \ldots, N_l]$ *with* $m < n_1 < \ldots < N_l \leq N$, *then* $S\big(1 : m, n_1 : N_l\big) = O_p(\delta_{N,T}^{-2})$ *for any constant* $l \geq 0$.

Proof. The result follows directly from Lemma B.5 and Lemma B.6. □

Lemma B.4 considers the case where none of the selected variables belong to the IOFs. In the following Lemma, we consider the case where only part of IOFs are selected. Without loss of generality, we assume that among the m IOFs, the kth to the last IOFs are selected.

Lemma B.9. $S\big(k : m, n_1 : n_l\big) \geq S\big(2 : m, n_1 : n_l\big)$ *for* $1 < k < m$, *and*

$$\textit{plim}\ S\big(2 : m, n_1 : n_l\big) > 0.$$

Proof. The first part follows directly from Lemma B.5. For the second part, let $y_t = (X'_{2:m,t}, X'_{n_1:n_l,t})'$, and $y = (y_1, \ldots, y_T)'$. Recall that:

$$S\big(2 : m, n_1 : n_l\big) = \sum_{j=1}^{r} S_j\big(2 : m, n_1 : n_l\big)$$

where $S_j\big(2 : m, n_1 : n_l\big) = \frac{1}{T}\sum_{t=1}^{T}(\tilde{f}_{jt} - \hat{A}'_j y_t)^2$. Then by Lemma B.5 we have:

$$S_j\big(2 : m, n_1 : n_l\big) = S_j\big(1 : m, n_1 : n_l\big) + b^2 T^{-1}\sum_{t=1}^{T}\hat{x}_{1t}^2,$$

where $\hat{x}_{1t}$ are the residuals in the regression of x_{1t} on y_t, and b is the coefficient of x_{1t} in the OLS regression of $\tilde{f}_{jt}$ on x_{1t} and y_t.

By (B.6) we have

$$\tilde{f}_t = HBX_{1:m,t} + o_p(1) = AX_{1:m,t} + o_p(1).$$

If we write $A = (A_1, \ldots, A_m)$, then $\|A_k\|^2 > 0$ for $k = 1, \ldots, m$ by Assumption 15(i) and the fact that H is nonsingular (Bai and Ng 2002). In the vector A_1 there must exit an element $a_{1j} \neq 0$. Thus we can write

$$\tilde{f}_{jt} = a_{1j}x_{1t} + Cy_t + o_p(1),$$

where $C = [a_{2j}, \ldots, a_{mj}, 0 \ldots, 0]$. It follows that $b^2 \xrightarrow{p} a_{1j}^2 > 0$.

Finally, we prove that plim $T^{-1}\sum_{t=1}^{T}\hat{x}_{1t}^2 > 0$ by contradiction. Suppose plim $T^{-1}\sum_{t=1}^{T}\hat{x}_{1t}^2 = 0$, define $z_t = (X'_{1:m,t}, X'_{n_1:n_l,t})'$, and write $x_{1t} = \hat{d}'y_t + \hat{x}_{1t}$, where $\hat{d}$ is the OLS estimator. Then:

$$\begin{aligned}
T^{-1}\sum_{t=1}^{T} z_t z_t' &= \begin{pmatrix} T^{-1}\sum_{t=1}^{T} x_{1t}^2 & T^{-1}\sum_{t=1}^{T}\hat{d}'y_t y_t' \\ T^{-1}\sum_{t=1}^{T} y_t y_t'\hat{d} & T^{-1}\sum_{t=1}^{T} y_t y_t' \end{pmatrix} \\
&= \begin{pmatrix} \hat{d}'\big(T^{-1}\sum_{t=1}^{T} y_t y_t'\big)\hat{d}' + \big(T^{-1}\sum_{t=1}^{T}\hat{x}_{1t}^2\big) & \hat{d}'\big(T^{-1}\sum_{t=1}^{T} y_t y_t'\big) \\ \big(T^{-1}\sum_{t=1}^{T} y_t y_t'\big)\hat{d} & T^{-1}\sum_{t=1}^{T} y_t y_t' \end{pmatrix} \\
&\xrightarrow{p} \begin{pmatrix} d'\Sigma_Y d & d'\Sigma_Y \\ \Sigma_Y d & \Sigma_Y \end{pmatrix}.
\end{aligned}$$

The last matrix is a singular matrix, which is a contradiction with Assumption 15(ii). Therefore we have:

$$\begin{aligned}\text{plim } S_j\big(2:m,n_1:n_l\big) &= \text{plim } S_j\big(1:m,n_1:n_l\big)+\text{plim }\big(b^2T^{-1}\sum_{t=1}^{T}\hat{x}_{1t}^2\big)\\ &= 0+a_{1j}^2\text{plim }\big(T^{-1}\sum_{t=1}^{T}\hat{x}_{1t}^2\big)>0.\end{aligned}$$

And thus plim $S\big(2:m,n_1:n_l\big)\geq$ plim $S_j\big(2:m,n_1:n_l\big)>0$.

□

Proof of Theorem 2.3

Proof. Since

$$\begin{aligned}&\mathbb{P}[\hat{m}=m,(\hat{n}_1,\ldots,\hat{N}_{\hat{m}})=(1,\ldots,m)]\\ =\;&\mathbb{P}[\hat{m}=m]\cdot\mathbb{P}[(\hat{n}_1,\ldots,\hat{N}_{\hat{m}})=(1,\ldots,m)|\hat{m}=m],\end{aligned}$$

and it is obvious that $\mathbb{P}[(\hat{n}_1,\ldots,\hat{N}_{\hat{m}})=(1,\ldots,m)|\hat{m}=m]\to 1$ as $N,T\to\infty$ by Lemma B.4, B.6 and B.8, it suffices to prove that $\mathbb{P}[\hat{m}=m]\to 1$.

When $l<m$:

$$\begin{aligned}&\mathbb{P}[\hat{m}=l]\\ =\;&\mathbb{P}[\min S(n_1:n_l)+l\cdot p(N,T)>\min S(n_1:n_m)+m\cdot p(N,T)]\\ =\;&\mathbb{P}[\min S(n_1:n_l)-\min S(n_1:n_m)>(m-l)\cdot p(N,T)]\end{aligned}$$

By Lemma B.4 and B.8, we have $\operatorname{plim\,inf} S(n_1:n_l)=\tau_l>0$, and $\operatorname{plim}\min S(n_1:n_l)=0$. Then we have $\mathbb{P}[\hat{m}=l]\to 0$ because $p(N,T)\to 0$.

Similarly, when $l>m$:

$$\begin{aligned}&\mathbb{P}[\hat{m}=l]\\ =\;&\mathbb{P}[\min S(n_1:n_l)+l\cdot p(N,T)<\min S(n_1:n_m)+m\cdot p(N,T)]\\ =\;&\mathbb{P}[\min S(n_1:n_m)-\min S(n_1:n_l)>(l-m)\cdot p(N,T)]\end{aligned}$$

From Lemma B.6, B.7 and B.8 we know that $\min S(n_1:n_m)-\min S(n_1:n_l)=O_p(\delta_{N,T}^{-2})$. By the assumption that $\delta_{N,T}^{-2}p(N,T)\to\infty$ as $N,T\to\infty$, $\cdot p(N,T)$ goes to zero slower than $\min S(n_1:n_m)-\min S(n_1:n_l)$, therefore $\mathbb{P}[\hat{m}=l]\to 0$ as $N,T\to\infty$. The desired result then follows easily. □

B.3 Proof of Theorem 2.4

Lemma B.10. *Define $\tilde{u}_k=[\tilde{u}_{k1},\ldots,\tilde{u}_{kT}]'$, then under Assumptions 11 to 13, $T^{-1/2}||\tilde{u}_k||=O_p(\delta_{N,T}^{-1})$.*

Proof. Since

$$T^{-1/2}\|\tilde{u}_k\| \leq T^{-1/2}\Big(\big\|\tilde{v}_k\big\| + \big\|(H_{NT}^k - H_0^k)f_t\big\|\Big)$$

where H^k denotes the kth row of H, and

$$T^{-1/2}\big\|\tilde{v}_k\big\| = \sqrt{\frac{1}{T}\sum_t^T \big(\tilde{f}_{kt} - H_{NT}^k f_t\big)^2} = O_p(\delta_{N,T}^{-1})$$

by Theorem 1 of Bai and Ng (2002), and

$$T^{-1/2}\big\|(H_{NT}^k - H_0^k)f_t\big\| = \sqrt{(H_{NT}^k - H_0^k)\Big(\frac{1}{T}\sum_t^T f_t f_t'\Big)(H_{NT}^k - H_0^k)'} = O_p(\delta_{N,T}^{-1})$$

because $\frac{1}{T}\sum_t^T f_t f_t' = O_p(1)$ by Assumption 11 and $H_{NT} - H_0 = O_p(\delta_{N,T}^{-1})$ under Assumptions 11 to 13 (Lemma 6 of Han and Inoue 2012). The result follows.

□

Lemma B.11. *Let* $\tilde{s}_{\mathcal{M}} = \big(|\tilde{\beta}_j|^{-1}\operatorname{sign}(\beta_{0j}), j \in \mathcal{M}\big)$, *and* $s_{\mathcal{M}} = \big(|\theta_j|^{-1}\operatorname{sign}(\beta_{0j}), j \in \mathcal{M}\big)$. *Under Assumption AL1,* $\big\|\tilde{s}_{\mathcal{M}}\big\| = O_p(1)$, *and*

$$\max_{j\notin\mathcal{M}}\Big\||\tilde{\beta}_j|\tilde{s}_{\mathcal{M}} - |\theta_j|s_{\mathcal{M}}\Big\| = o_p(1).$$

Proof. First, we have

$$\begin{aligned}\max_{j\in\mathcal{M}}\Big||\tilde{\beta}_j|/|\theta_j| - 1\Big| &\leq \sum_{j\in\mathcal{M}}\frac{\big||\tilde{\beta}_j| - |\theta_j|\big|}{|\theta_j|} \leq \max_{j\in\mathcal{M}}\Big|\tilde{\beta}_j - \theta_j\Big|\sum_{j\in\mathcal{M}}\frac{1}{|\theta_j|}\\ &\leq O_p(r_{NT}^{-1}) = o_p(1)\end{aligned}$$

by Assumptions AL1. Then,

$$\begin{aligned}\big\|\tilde{s}_{\mathcal{M}}\big\| &\leq \big\|\tilde{s}_{\mathcal{M}} - s_{\mathcal{M}}\big\| + \big\|s_{\mathcal{M}}\big\| = \sqrt{\sum_{j\in\mathcal{M}}\Big|\frac{1}{|\tilde{\beta}_j|} - \frac{1}{|\theta_i|}\Big|^2} + \sqrt{\sum_{j\in\mathcal{M}}\frac{1}{|\theta_i|^2}}\\ &\leq \Big(\max_{j\in\mathcal{M}}\Big||\theta_j|/|\tilde{\beta}_j| - 1\Big| + 1\Big)\sqrt{\sum_{j\in\mathcal{M}}\frac{1}{|\theta_i|^2}} = O_p(1).\end{aligned}$$

For the second part of the lemma, notice that

$$\Big\||\tilde{\beta}_j|\tilde{s}_{\mathcal{M}} - |\theta_j|s_{\mathcal{M}}\Big\| = \Big\||\tilde{\beta}_j|\tilde{s}_{\mathcal{M}} - |\theta_j|\tilde{s}_{\mathcal{M}} + |\theta_j|\tilde{s}_{\mathcal{M}} - |\theta_j|s_{\mathcal{M}}\Big\| \leq \Big\||\tilde{\beta}_j|\tilde{s}_{\mathcal{M}} - |\theta_j|\tilde{s}_{\mathcal{M}}\Big\| + \Big\||\theta_j|\tilde{s}_{\mathcal{M}} - |\theta_j|s_{\mathcal{M}}\Big\|.$$

First,

$$\begin{aligned}\max_{j\notin\mathcal{M}}\left\||\theta_j|\tilde{s}_{\mathcal{M}}-|\theta_j|s_{\mathcal{M}}\right\| &\leq \max_{j\notin\mathcal{M}}|\theta_j|\cdot\left\|\tilde{s}_{\mathcal{M}}-s_{\mathcal{M}}\right\| \leq C_2\sqrt{\sum_{j\in\mathcal{M}}\left|\frac{|\tilde{\beta}_j|-|\theta_i|}{|\tilde{\beta}_j||\theta_i|}\right|^2}\\ &\leq \max_{j\in\mathcal{M}}\left|\tilde{\beta}_j-\theta_j\right|\sqrt{\sum_{j\in\mathcal{M}}\frac{C_2^2}{|\tilde{\beta}_j|^2|\theta_i|^2}} \leq O_p(r_{NT}^{-1})\sqrt{\sum_{j\in\mathcal{M}}\left(\frac{C_2}{|\theta_i|^2}\right)^2\left(\frac{\tilde{\beta}_j}{|\theta_i|}\right)^2}\\ &= O_p(r_{NT}^{-1})=o_p(1)\end{aligned}$$

by Assumption AL1. Second,

$$\max_{j\notin\mathcal{M}}\left\||\tilde{\beta}_j|\tilde{s}_{\mathcal{M}}-|\theta_j|\tilde{s}_{\mathcal{M}}\right\| \leq \max_{j\notin\mathcal{M}}\left|\tilde{\beta}_j-\theta_j\right|\cdot\left\|\tilde{s}_{\mathcal{M}}\right\| = O_p(r_{NT}^{-1})=o_p(1).$$

Then the result follows.

□

Proof of Theorem 2.4

Proof. The first order conditions for the Adaptive Lasso implies that $\hat{\beta}^L=[\hat{\beta}_1^L,\ldots,\hat{\beta}_{m+p}^L]'$ is the unique solution if

$$Z_j'(\tilde{f}_k-Z\hat{\beta}^L) = \lambda_{NT}w_j\,\mathrm{sign}(\hat{\beta}_j^L) \text{ for } \hat{\beta}_j^L\neq 0 \tag{B.7}$$

$$|Z_j'(\tilde{f}_k-Z\hat{\beta})| < \lambda_{NT}w_j \text{ for } \hat{\beta}_j=0 \tag{B.8}$$

and the vectors in $Z_{\mathcal{P}}$ are linearly independent, where $Z_j=[z_{j1},\ldots,z_{jT}]'$, $\tilde{f}=[\tilde{f}_{k1},\ldots,\tilde{f}_{kT}]'$, $Z=[Z_1,Z_2,\ldots,Z_{m+p}]$, $Z_{\mathcal{P}}=[Z_{\mathcal{P}1},\ldots,Z_{\mathcal{P}p}]$. Define $\tilde{s}_{\mathcal{M}}=[w_j\,\mathrm{sign}(\beta_{0j}),j\in\mathcal{M}]$, and

$$\hat{\beta}_{\mathcal{M}}=(Z_{\mathcal{M}'}Z_{\mathcal{M}})^{-1}(Z_{\mathcal{M}}'\tilde{f}_k-\lambda_{NT}\tilde{s}_{\mathcal{M}}). \tag{B.9}$$

Since $\tilde{f}_k=Z_{\mathcal{M}}\beta_{\mathcal{M}0}+\tilde{u}_{kt}$, then

$$\hat{\beta}_{\mathcal{M}}=\beta_{\mathcal{M}0}+T^{-1}\Sigma_{\mathcal{M}T}^{-1}(Z_{\mathcal{M}}'\tilde{u}_k-\lambda_{NT}\tilde{s}_{\mathcal{M}}). \tag{B.10}$$

Let $\hat{\beta}=[\hat{\beta}_{\mathcal{M}}'\quad\mathbf{0}_p']'$, where $\mathbf{0}_p$ is a $p\times 1$ vector of zeros, then it follows that $\hat{\beta}=_s\beta_0$ if

$$\hat{\beta}_{\mathcal{M}}=_s\beta_{\mathcal{M}0}. \tag{B.11}$$

Moreover, we have $\hat{\beta}^L=\hat{\beta}$ if (B.11) holds and

$$|Z_j'(\tilde{f}_k-Z_{\mathcal{M}}\hat{\beta}_{\mathcal{M}})|<\lambda_{NT}w_j \text{ for } j\notin\mathcal{M}. \tag{B.12}$$

Thus, $\hat{\beta}^L=_s\beta_0$ if (B.11) and (B.12) hold.

From (B.10) we have $\tilde{f}_k - Z_{\mathcal{M}}\hat{\beta}_{\mathcal{M}} = \tilde{u}_k - Z_{\mathcal{M}}(\hat{\beta}_{\mathcal{M}} - \beta_{\mathcal{M}0}) = D_T\tilde{u}_k + Z_{\mathcal{M}}\Sigma_{\mathcal{M}T}^{-1}\tilde{s}_{\mathcal{M}}\lambda_{NT}/T$, where $D_T = I_T - Z_{\mathcal{M}}\Sigma_{\mathcal{M}T}^{-1}Z_{\mathcal{M}}'/T$. Then it follows that $\hat{\beta}^L =_s \beta_0$ if

$$\operatorname{sign}(\beta_{0j})(\beta_{0j} - \hat{\beta}_j) < |\beta_{0j}| \text{ for } j \in \mathcal{M}, \tag{B.13}$$

$$\left|Z_j'\left(D_T\tilde{u}_k + Z_{\mathcal{M}}\Sigma_{\mathcal{M}T}^{-1}\tilde{s}_{\mathcal{M}}\lambda_{NT}/T\right)\right| < \lambda_{NT}w_j \text{ for } j \notin \mathcal{M}. \tag{B.14}$$

Thus,

$$\begin{aligned}
\mathbb{P}\Big\{\hat{\beta}^L \neq_s \beta_0\Big\} &\leq \mathbb{P}\Big\{T^{-1}\big|e_j'\Sigma_{\mathcal{M}T}^{-1}Z_{\mathcal{M}}\tilde{u}_k\big| \geq |\beta_{0j}|/2 \text{ for some } j \in \mathcal{M}\Big\} \\
&+ \mathbb{P}\Big\{\big|e_j'\Sigma_{\mathcal{M}T}^{-1}\tilde{s}_{\mathcal{M}}\big|\lambda_{NT}/T \geq |\beta_{0j}|/2 \text{ for some } j \in \mathcal{M}\Big\} \\
&+ \mathbb{P}\Big\{\big|Z_j'D_T\tilde{u}_k\big| \geq (1-\kappa-\epsilon)\lambda_{NT}w_j \text{ for some } j \notin \mathcal{M}\Big\} \\
&+ \mathbb{P}\Big\{T^{-1}\big|Z_j'Z_{\mathcal{M}}\Sigma_{\mathcal{M}T}^{-1}\tilde{s}_{\mathcal{M}}\big| \geq (\kappa+\epsilon)w_j \text{ for some } j \notin \mathcal{M}\Big\} \\
&= \mathbb{P}\{I\} + \mathbb{P}\{II\} + \mathbb{P}\{III\} + \mathbb{P}\{IV\},
\end{aligned}$$

for any $0 < \kappa < \kappa + \epsilon < 1$, where $e_j' = [0, \ldots, 0, 1, 0, \ldots, 0]$ is the vector that selects the jth element.

First, since $T^{-1/2}\big\|e_j'\Sigma_{\mathcal{M}T}^{-1}Z_{\mathcal{M}}\big\| = \sqrt{\big|e_j'\Sigma_{\mathcal{M}T}^{-1}e_j\big|} \leq \sqrt{\operatorname{Tr}(\Sigma_{\mathcal{M}T}^{-1})} \leq \sqrt{m/\tau_1}$ *a.s* for $j \in \mathcal{M}$,

$$\begin{aligned}
\mathbb{P}\{I\} &= \mathbb{P}\Big\{T^{-1}\big|e_j'\Sigma_{\mathcal{M}T}^{-1}Z_{\mathcal{M}}\tilde{u}_k\big| \geq |\beta_{0j}|/2 \text{ for some } j \in \mathcal{M}\Big\} \\
&\leq \mathbb{P}\Big\{T^{-1}\big\|e_j'\Sigma_{\mathcal{M}T}^{-1}Z_{\mathcal{M}}\big\| \cdot \big\|\tilde{u}_k\big\| \geq |\beta_{0j}|/2 \text{ for some } j \in \mathcal{M}\Big\} \\
&\leq \mathbb{P}\Big\{T^{-1/2}\big\|\tilde{u}_k\big\| \geq \sqrt{\tau_1/m}|\beta_{0j}|/2\Big\} \cdot m \to 0
\end{aligned}$$

because $T^{-1/2}\big\|\tilde{u}_k\big\|$ is $o_p(1)$ by Lemma B.10.

Second, since $\big\|e_j'\Sigma_{\mathcal{M}T}^{-1}\big\| \leq \sqrt{\operatorname{Tr}\big(\Sigma_{\mathcal{M}T}^{-2}\big)} \leq \sqrt{m}/\tau_1$ *a.s*,

$$\big|e_j'\Sigma_{\mathcal{M}T}^{-1}\tilde{s}_{\mathcal{M}}\big|\lambda_{NT}/T \leq \big\|e_j'\Sigma_{\mathcal{M}T}^{-1}\big\|\big\|\tilde{s}_{\mathcal{M}}\big\|\lambda_{NT}/T = O_p(\frac{\lambda_{NT}}{T\tau_1}) = o_p(1)$$

by Assumption AL2, then $\mathbb{P}\{II\} \to 0$.

Third, $\big\|Z_j'D_T\big\| = \big|Z_j'D_TZ_j\big|^{1/2} \leq \big|Z_j'Z_j\big|^{1/2} = \sqrt{T}$, and $w_j^{-1} = \big|\tilde{\beta}_j\big| \leq |\theta_j| + |\tilde{\beta}_j - \theta_j| \leq C_2 + O_p(1/r_{NT})$ for $j \notin \mathcal{M}$, then for a large enough C, $\mathbb{P}\{w_j^{-1} \leq C(C_2 + 1/r_{NT})\} \to 1$. Therefore,

$$\begin{aligned}
\mathbb{P}\{III\} &\leq \mathbb{P}\Big\{\big|Z_j'D_T\tilde{u}_k\big| \geq \frac{(1-\kappa-\epsilon)\lambda_{NT}}{C(C_2+1/r_{NT})} \text{ for some } j \notin \mathcal{M}\Big\} + o(1) \\
&= \mathbb{P}\Big\{\max_{j\notin\mathcal{M}}\big|Z_j'D_T\tilde{u}_k\big| \geq \frac{(1-\kappa-\epsilon)\lambda_{NT}}{C(C_2+1/r_{NT})}\Big\} + o(1) \\
&\leq \mathbb{P}\Big\{\frac{\delta_{N,T}}{\sqrt{T}}\big\|\tilde{u}_k\big\| \geq \frac{(1-\kappa-\epsilon)\lambda_{NT}\delta_{N,T}}{C(C_2+1/r_{NT})T}\Big\} + o(1) \to 0
\end{aligned}$$

by Assumption AL2 and Lemma B.10.

Finally, $\big\|Z_j'Z_{\mathcal{M}}\Sigma_{\mathcal{M}T}^{-1}\big\|T^{-1} \leq \big\|Z_j\big\|T^{-1/2} \cdot \big\|Z_{\mathcal{M}}\Sigma_{\mathcal{M}T}^{-1}\big\|T^{-1/2}$, since $\big\|Z_j\big\|T^{-1/2} = 1$ by construction, and

$$\big\|Z_{\mathcal{M}}\Sigma_{\mathcal{M}T}^{-1}\big\|T^{-1/2} = \sqrt{\operatorname{Tr}\big(\Sigma_{\mathcal{M}T}^{-1}\big)} \leq (m/\tau_1)^{1/2} \text{ a.s},$$

thus $\left\|Z_j'Z_{\mathcal{M}}\Sigma_{\mathcal{M}T}^{-1}\right\|T^{-1} \leq (m/\tau_1)^{1/2}$ *a.s.* Therefore,

$$\begin{aligned}
& \max_{j\notin\mathcal{M}}\left(T^{-1}\left|Z_j'Z_{\mathcal{M}}\Sigma_{\mathcal{M}T}^{-1}\tilde{s}_{\mathcal{M}}w_j^{-1}\right| - T^{-1}\left|Z_j'Z_{\mathcal{M}}\Sigma_{\mathcal{M}T}^{-1}s_{\mathcal{M}}\theta_j\right|\right) \\
\leq & \max_{j\notin\mathcal{M}}\left(T^{-1}\left|Z_j'Z_{\mathcal{M}}\Sigma_{\mathcal{M}T}^{-1}\left(\tilde{s}_{\mathcal{M}}w_j^{-1} - s_{\mathcal{M}}\theta_j\right)\right|\right) \\
\leq & \max_{j\notin\mathcal{M}}\left(T^{-1}\left\|Z_j'Z_{\mathcal{M}}\Sigma_{\mathcal{M}T}^{-1}\right\|\right)\left\||\tilde{\beta}_j|\tilde{s}_{\mathcal{M}} - |\theta_j|s_{\mathcal{M}}\right\| = o_p(1)
\end{aligned}$$

by Lemma B.11. Moreover, $T^{-1}\left|Z_j'Z_{\mathcal{M}}\Sigma_{\mathcal{M}T}^{-1}s_{\mathcal{M}}\theta_j\right| \leq \kappa$ by Assumption AL3, it follows that

$$\begin{aligned}
\mathbb{P}\{IV\} &= \mathbb{P}\left\{T^{-1}\left|Z_j'Z_{\mathcal{M}}\Sigma_{\mathcal{M}T}^{-1}\tilde{s}_{\mathcal{M}}w_j^{-1}\right| \geq \kappa + \epsilon \text{ for some } j\notin\mathcal{M}\right\}. \\
&\leq \mathbb{P}\left\{\max_{j\notin\mathcal{M}}\left(T^{-1}\left|Z_j'Z_{\mathcal{M}}\Sigma_{\mathcal{M}T}^{-1}\tilde{s}_{\mathcal{M}}w_j^{-1}\right| - T^{-1}\left|Z_j'Z_{\mathcal{M}}\Sigma_{\mathcal{M}T}^{-1}s_{\mathcal{M}}\theta_j\right|\right) \geq \epsilon\right\} \\
&+ \mathbb{P}\left\{T^{-1}\left|Z_j'Z_{\mathcal{M}}\Sigma_{\mathcal{M}}^{-1}s_{\mathcal{M}}\theta_j\right| \geq \kappa \text{ for some } j\notin\mathcal{M}\right\} \to 0
\end{aligned}$$

by Assumption AL3. This completes the proof. □

B.4 Proof of Theorem 2.6

From the definition of eigenvectors and eigenvalues we have $(1/NT)XX'\tilde{f} = \tilde{f}V_{NT}$. Note that

$$XX' = f\Lambda'\Lambda f' + f\Lambda'e' + e\Lambda f' + ee'$$

where $e = [e_1, \ldots, e_t]'$. It follows that

$$\tilde{V}_t = \tilde{f}_t - Hf_t = V_{NT}^{-1}\left(\frac{1}{T}\sum_{s=1}^{T}\tilde{f}_s\gamma_N(s,t) + \frac{1}{T}\sum_{s=1}^{T}\tilde{f}_s\zeta_{st} + \frac{1}{T}\sum_{s=1}^{T}\tilde{f}_s\eta_{st} + \frac{1}{T}\sum_{s=1}^{T}\tilde{f}_s\xi_{st},\right) \tag{B.15}$$

where $\gamma_N(s,t) = 1/N\sum_{i=1}^{N}E(e_{is}e_{it})$, $\zeta_{st} = 1/N\sum_{i=1}^{N}\left(e_{is}e_{it} - E(e_{is}e_{it})\right), \eta_{st} = f_s'\Lambda'e_t/N$, and $\xi_{st} = f_t'\Lambda'e_s/N$. Note that under Assumption 16, $\gamma_N(s,t) = 0$ for $s \neq t$, and $\gamma_N(t,t) = \gamma_N = 1/N\sum_{i=1}^{N}\sigma_{ei}^2 = O(1)$. Then Theorem 2.6 follows easily from the following 4 lemmas:

Lemma B.12. *Under Assumptions 11, 12, 13 and 16, we have*

$$V_{NT}^{-1}\frac{\sqrt{N}}{T^{3/2}}\sum_{t=1}^{T}\sum_{s=1}^{T}\tilde{f}_s\gamma_N(s,t) = o_p(1).$$

Proof. First, since $V_{NT} \xrightarrow{p} V > 0$, $\|V_{NT}^{-1}\| = O_p(1)$. Further,

$$\frac{\sqrt{N}}{T^{3/2}}\sum_{t=1}^{T}\sum_{s=1}^{T}\tilde{f}_s\gamma_N(s,t) = \frac{\sqrt{N}}{T^{3/2}}\sum_{s=1}^{T}\left(\left(\tilde{f}_s - Hf_s\right)\sum_{t=1}^{T}\gamma_N(s,t)\right) + H\frac{\sqrt{N}}{T^{3/2}}\sum_{s=1}^{T}\left(f_s\sum_{t=1}^{T}\gamma_N(s,t)\right). \tag{B.16}$$

For the first the term on the right, we have

$$\frac{\sqrt{N}}{T^{3/2}}\left\|\sum_{s=1}^{T}\left(\left(\tilde{f}_s - Hf_s\right)\sum_{t=1}^{T}\gamma_N(s,t)\right)\right\| \leq \frac{\sqrt{N}}{\sqrt{T}}\sqrt{\frac{1}{T}\sum_{s=1}^{T}\left\|\tilde{f}_s - Hf_s\right\|^2}\sqrt{\frac{1}{T}\sum_{s=1}^{T}\left(\sum_{t=1}^{T}\gamma_N(s,t)\right)^2}.$$

On the righthand side of the inequality, both $\sqrt{N/T}$ and the last term are $O(1)$ by Assumption 16, and the middle term is $O_p(\delta_{N,T}^{-1})$ by Bai and Ng (2002), where $\delta_{N,T} = min[\sqrt{N}, \sqrt{T}]$. Next, for the second term on the righthand side of (B.16), we have

$$H\frac{\sqrt{N}}{T^{3/2}}\sum_{s=1}^{T}\left(f_s\sum_{t=1}^{T}\gamma_N(s,t)\right) = \gamma_N H\frac{\sqrt{N}}{T}\left(\frac{1}{\sqrt{T}}\sum_{s=1}^{T}f_s\right) = \frac{\sqrt{N}}{T}O_p(1) = o_p(1),$$

because $\gamma_N = O(1)$, $\sqrt{N}/T = o(1)$ by Assumption 16, $\|H\|$ is $O_p(1)$ since $H \xrightarrow{p} H_0 > 0$, and $1/\sqrt{T}\sum_{s=1}^{T} f_s$ is $O_p(1)$ by Assumption 16. □

Lemma B.13. *Under Assumptions 11, 12, 13 and 16, we have*

$$V_{NT}^{-1}\frac{\sqrt{N}}{T^{3/2}}\sum_{t=1}^{T}\sum_{s=1}^{T}\tilde{f}_s\zeta_{st} = o_p(1).$$

Proof. Similarly, we can write

$$N^{1/2}T^{-3/2}\sum_{t=1}^{T}\sum_{s=1}^{T}\tilde{f}_s\zeta_{st} = N^{1/2}T^{-3/2}\sum_{t=1}^{T}\sum_{s=1}^{T}\left(\tilde{f}_s - Hf_s\right)\zeta_{st} + N^{1/2}T^{-3/2}H\sum_{t=1}^{T}\sum_{s=1}^{T}f_s\zeta_{st}. \quad \text{(B.17)}$$

For the first term on the righthand side of (B.17), we have

$$N^{1/2}T^{-3/2}\left\|\sum_{t=1}^{T}\sum_{s=1}^{T}\left(\tilde{f}_s - Hf_s\right)\zeta_{st}\right\| \leq N^{1/2}T^{-3/2}\sum_{s=1}^{T}\left(\left\|\tilde{f}_s - Hf_s\right\|\left|\sum_{t=1}^{T}\zeta_{st}\right|\right)$$

$$\leq \sqrt{\frac{1}{T}\sum_{s=1}^{T}\left\|\tilde{f}_s - Hf_s\right\|^2}\sqrt{\frac{1}{T}\sum_{s=1}^{T}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\sqrt{N}\zeta_{st}\right)^2} = O_p(\delta_{N,T}^{-1}),$$

since $1/T\sum_{s=1}^{T}\left\|\tilde{f}_s - Hf_s\right\|^2$ is $O_p(\delta_{N,T}^{-2})$, and

$$\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\sqrt{N}\zeta_{st}\right|^2 = \left|\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(e_{it}e_{is} - E(e_{it}e_{is})\right)\right|^2 = O_p(1)$$

by Assumption 16. For the second term on the righthand side of (B.17), we have

$$N^{1/2}T^{-3/2}H\sum_{t=1}^{T}\sum_{s=1}^{T}f_s\zeta_{st} = \frac{1}{\sqrt{T}}\left(\frac{1}{T\sqrt{N}}\sum_{s=1}^{T}\sum_{t=1}^{T}\sum_{i=1}^{N}f_s\left(e_{it}e_{is} - E(e_{it}e_{is})\right)\right) = O_p(T^{-1/2})$$

by Assumption 16. The result then follows since $\|V_{NT}^{-1}\| = O_p(1)$.

□

Lemma B.14. *Under Assumptions 11, 12, 13 and 16, we have*

$$V_{NT}^{-1}\frac{\sqrt{N}}{T^{-3/2}}\sum_{t=1}^{T}\sum_{s=1}^{T}\tilde{f}_s\eta_{st} \xrightarrow{d} N(0,\Xi)$$

as N and T go to infinity, where $\Xi = V^{-1}Q\Psi Q'V^{-1}$.

Proof. By definition of η_{st} we have

$$V_{NT}^{-1}\frac{\sqrt{N}}{T^{-3/2}}\sum_{t=1}^{T}\sum_{s=1}^{T}\tilde{f}_s\eta_{st} = V_{NT}^{-1}\left(\frac{1}{T}\sum_{s=1}^{T}\tilde{f}_s f_s'\right)\left(\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{i=1}^{N}\lambda_i e_{it}\right),$$

the result follows from Assumption 16 and the facts that $V_{N,T} \xrightarrow{p} V$ (Stock and Watson 2002)and $1/T\sum_{s=1}^{T}\tilde{f}_s f_s' \xrightarrow{p} Q$ (Lemma B.1). □

Lemma B.15. *Under Assumptions 11, 12, 13 and 16, we have*

$$V_{NT}^{-1}\frac{\sqrt{N}}{T^{3/2}}\sum_{t=1}^{T}\sum_{s=1}^{T}\tilde{f}_s\xi_{st} = o_p(1).$$

Proof. First, by definition of ξ_{st}, we have

$$\frac{\sqrt{N}}{T^{3/2}}\sum_{t=1}^{T}\sum_{s=1}^{T}\tilde{f}_s\xi_{st} = \frac{\sqrt{N}}{T^{3/2}}\sum_{t=1}^{T}\sum_{s=1}^{T}\tilde{f}_s e_s'\Lambda f_t/N = \left(\frac{1}{T}\sum_{s=1}^{T}\tilde{f}_s e_s'\Lambda/\sqrt{N}\right)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}f_t\right).$$

Notice that $\left\|1/\sqrt{T}\sum_{t=1}^{T}f_t\right\| = O_p(1)$ by assumption, and

$$\frac{1}{T}\sum_{s=1}^{T}\tilde{f}_s e_s'\Lambda/\sqrt{N} = \frac{1}{T}\sum_{s=1}^{T}\left(\tilde{f}_s - Hf_s\right)e_s'\Lambda/\sqrt{N} + H\frac{1}{T}\sum_{t=1}^{T}f_s e_s'\Lambda/\sqrt{N}.$$

For the first term, we have

$$\left\|\frac{1}{T}\sum_{s=1}^{T}\left(\tilde{f}_s - Hf_s\right)e_s'\Lambda/\sqrt{N}\right\| \leq \sqrt{\frac{1}{T}\sum_{s=1}^{T}\left\|\tilde{f}_s - Hf_s\right\|^2}\sqrt{\frac{1}{T}\sum_{s=1}^{T}\left\|e_s'\Lambda/\sqrt{N}\right\|^2} = O_p(\delta_{N,T}^{-1})$$

by Assumption 16. The second term is $O_p(T^{-1/2})$ because

$$\left\|H\frac{1}{T}\sum_{t=1}^{T}f_s e_s'\Lambda/\sqrt{N}\right\| \leq T^{-1/2}\|H\|\left\|\frac{1}{\sqrt{NT}}\sum_{s=1}^{T}\sum_{i=1}^{N}f_s\lambda_i' e_s\right\| = O_p(T^{-1/2})$$

by Assumption 16. Therefore we have

$$\left\|\frac{1}{T}\sum_{s=1}^{T}\tilde{f}_s e_s'\Lambda/\sqrt{N}\right\| = o_p(1)$$

and the result follows. □

As a result, the limit distribution of $1/\sqrt{T}\sum_{t=1}^{T}\sqrt{N}\tilde{V}_t$ depends only on the corresponding sum of the second term in (B.15), whose distribution was derived in Lemma B.14. Then Theorem 2.6 follows easily.

B.5 Tables and Figures

Table B.1: Candidates for Observed Factors

	Short Name	Long Name	T code
1	IP	industrial production: total index	5
2	IPMFG	industrial production: manufacturing	5
3	IPXMCA	capacity util rate: manufacturing, total	1
4	LHELX	employment: ratio; help-wanted ads:no. unemployment clf	4
5	LHUR	unemployment rate: all workers, 16 years over	1
6	LPNAG	employment on nonag. payrolls: total	5
7	LEHCC	avg hr earnings of constr wkrs: construction	6
8	LEHM	avg hr earings of prod wrks: manufacturing	6
9	HSFR	housing starts: nonfarm (1947-58) ; total farm & nonfarm(1959-)	4
10	HSBR	housing authorized by build: total new priv housing units	4
11	MSMTQ	manufacturing & trade: total	5
12	MSMQ	manufacturing & trade: manufacturing; total	5
13	WTQ	merchant wholesalers: total	5
14	RTQ	retail trade:total	5
15	IVMTQ	manufacturing & trade inventories: total	5
16	PMI	purchasing managers' index	1
17	PMP	napm production index	1
18	PMNO	napm new orders index	1
19	PMNV	napm inventories index	1
20	PMEMP	napm employment index	1
21	MO	mfg new orders: all manufacturing industries, total	5
22	MDO	mfg new orders: durable goods industries, total	5
23	FM2	money stock: m2	6
24	FMFBA	monetary base, adj for reserve requirement changes	6
25	FSNCOM	NYSE common stock price index: composite	5
26	FSPCOM	S&P common stock price index: composite	5
27	FSPIN	S&P common stock price index: industries	5
28	FSPCAP	S&P common stock price index: capital goods	5
29	FYFF	interest rate: federal funds	2
30	FYCP90	interest rate: 90 day commercial paper	2
31	FYGM3	interest rate: U.S. treasury bills, sec mkt, 3-m0	2
32	FYGM6	interest rate: U.S. treasury bills, sec mkt, 3-m0	2
33	FYGT1	interest rate: U.S. treasury const maturities, 1-yr	2
34	FYGT5	interest rate: U.S. treasury const maturities, 5-yr	2
35	FYGT10	interest rate: U.S. treasury const maturities, 10-yr	2
36	FYAAAC	bond yield: moody's aaa corporate	2
37	FYBAAC	bond yield: moody's baa corporate	2
38	FYFHA	secondary market yields on fha mortgages	2
39	EXRUS	United States effective exchange rate	5
40	EXRGER	foreign exchange rate: Germany	5
41	EXRJAN	foreign exchange rate: Japan	5
42	EXRUK	foreign exchange rate: United Kingdom	5
43	EXRCAN	foreign exchange rate: Canada	5
44	PWFSA	producer price index: finished goods	6
45	PUNEW	cpi-u: all items	6
46	PUC	cpi-u: commodities	6
47	GMDC	pce, impl pr defl: pce	6
48	Market	Market minus risk free rate	1
49	SMB	small minus big	1
50	HML	high minus low	1

Bibliography

[1] Acemoglu, D., V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi (2012). The network origins of aggregate fluctuations. *Econometrica 80*(5), 1977–2016.

[2] Ahn, S. and A. Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*. Forthcoming.

[3] Altug, S. (1989). Time-to-build and aggregate fluctuations: some new evidence. *International Economic Review*, 889–920.

[4] Andrews, D. (1993a). Tests for parameter instability and structural change with unknown change point. *Econometrica 61*(4), 821–856.

[5] Andrews, D. (2003). Tests for parameter instability and structural change with unknown change point: a corrigendum. *Econometrica*, 395–397.

[6] Andrews, D. W. K. (1993b). Tests for parameter instability and structural change with unknown change point. *Econometrica 61*(4), pp. 821–856.

[7] Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*(1), 135–171.

[8] Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*(4), 1229–1279.

[9] Bai, J. and S. Ng (2002a). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

[10] Bai, J. and S. Ng (2002b). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

[11] Bai, J. and S. Ng (2006a). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica 74*(4), 1133–1150.

[12] Bai, J. and S. Ng (2006b). Evaluating latent and observed factors in macroeconomics and finance. *Journal of Econometrics 131*(1-2), 507–537.

[13] Bai, J. and S. Ng (2011). Principal components estimation and identification of the factors. *Columbia University, Working paper*.

[14] Bai, J. and P. Perron (1998). Estimating and testing linear models with multiple structural changes. *Econometrica 66*(1), 47–78.

[15] Bai, J. and P. Perron (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics 18*(1), 1–22.

[16] Banerjee, A., M. Marcellino, and I. Masten (2008). Forecasting macroeconomic variables using diffusion indexes in short samples with structural change. *CEPR. DP. 6706*.

[17] Bates, B., M. Plagborg-Moller, J. Stock, and M. Watson (2013). Consistent factor estimation in dynamic factor models with structural instability. *Journal of Econometrics*. Forthcoming.

[18] Bernanke, B., J. Boivin, and P. Eliasz (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach. *The Quarterly Journal of Economics 120*(1), 387.

[19] Bickel, P., Y. Ritov, and A. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

[20] Boivin, J. and M. Giannoni (2006). Dsge models in a data-rich environment. Technical report, National Bureau of Economic Research.

[21] Breitung, J. and S. Eickmeier (2011). Testing for structural breaks in dynamic factor models. *Journal of Econometrics 163*(1), 71–84.

[22] Brown, R., J. Durbin, and J. Evans (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society. Series B (Methodological)*, 149–192.

[23] Brown, S. (1989). The number of factors in security returns. *Journal of Finance*, 1247–1262.

[24] Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica 51*(5), 1281–1304.

[25] Charnavoki, V. and J. Dolado (2012). The effects of global shocks on small commodity-exporting economies: New evidence from Canada. *CEPR. DP. 8825*.

[26] Chen, B. and Y. Hong (2012). Testing for smooth structural changes in time series models via nonparametric regression. *Econometrica 80*(3), 1157–1183.

[27] Chen, L. (2013). Identifying observed factors in high dimensional factor models. *Universidad Carlos III de Madrid, Working paper*.

[28] Chen, L., J. Dolado, and J. Gonzalo (2013). Dectecing big structural breaks in large factor models. *Universidad Carlos III de Madrid, Working paper*.

[29] Chen, N., R. Roll, and S. Ross (1986). Economic forces and the stock market. *Journal of Business*, 383–403.

[30] Christiano, L. and T. Fitzgerald (1999). The band pass filter. Technical report, National Bureau of Economic Research.

[31] Cochrane, J. and M. Piazzesi (2005). Bond risk premia. *American Economic Review*, 138–160.

[32] Diebold, F. and C. Chen (1996). Testing structural stability with endogenous breakpoint a size comparison of analytic and bootstrap procedures. *Journal of Econometrics 70*(1), 221–241.

[33] Dupor, B. (1999). Aggregation and irrelevance in multi-sector models. *Journal of Monetary Economics 43*(2), 391–409.

[34] Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of statistics 32*(2), 407–499.

[35] Fama, E. and K. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics 33*(1), 3–56.

[36] Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

[37] Fernández-Villaverde, J., P. A. Guerrón-Quintana, and J. Rubio-Ramírez (2010). Reading the recent monetary history of the US, 1959-2007. *Federal Reserve Board of St. Louis Review, 92*, 1–28.

[38] Forni, M., D. Giannone, M. Lippi, and L. Reichlin (2009). Opening the black box: Structural factor models with large cross-sections. *Econometric Theory 25*(5), 1319–1347.

[39] Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics 82*(4), 540–554.

[40] Forni, M. and M. Lippi (2001). The generalized dynamic factor model: representation theory. *Econometric Theory 17*(06), 1113–1141.

[41] Forni, M. and L. Reichlin (1998). Let's get real: a factor analytical approach to disaggregated business cycle dynamics. *The Review of Economic Studies 65*(3), 453–473.

[42] Frank, I. and J. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 109–135.

[43] Gabaix, X. (2011). The granular origins of aggregate fluctuations. *Econometrica 79*(3), 733–772.

[44] Gagliardini, P. and C. Gouriéroux (2011). Efficiency in large dynamic panel models with common factor. *U. of Toronto (mimeo)*.

[45] Gali, J. and L. Gambetti (2009). On the sources of the great moderation. *American Economic Journal. Macroeconomics 1*(1), 26–57.

[46] Geweke, J. (1978). *The dynamic factor analysis of economic time series models*. Social Systems Research Institute, University of Wisconsin-Madison.

[47] Giannone, D. (2007). Discussion on: "Forecasting in dynamic factor models subject to structural instability" by Stock and Watson. In *New Developments in Dynamic Factor Modelling*. Bank of England, Centre for Central Banking Studies.

[48] Goyal, A. and P. Santa-Clara (2003). Idiosyncratic risk matters! *The Journal of Finance 58*(3), 975–1008.

[49] Han, X. and A. Inoue (2012). Tests for parameter instability in dynamic factor models. *North Carolina State University, Working paper*.

[50] Hansen, B. (2000). Testing for structural change in conditional models. *Journal of Econometrics 97*(1), 93–115.

[51] Harding, M. (2008). Explaining the single factor bias of arbitrage pricing models in finite samples. *Economics Letters 99*(1), 85–88.

[52] Horvath, M. (1998). Cyclicality and sectoral linkages: Aggregate fluctuations from independent sectoral shocks. *Review of Economic Dynamics 1*(4), 781–808.

[53] Horvath, M. (2000). Sectoral shocks and aggregate fluctuations. *Journal of Monetary Economics 45*(1), 69–106.

[54] Huang, J., S. Ma, and C. Zhang (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica 18*(4), 1603.

[55] Kim, D. and P. Perron (2009). Unit root tests allowing for a break in the trend function at an unknown time under both the null and alternative hypotheses. *Journal of Econometrics 148*(1), 1–13.

[56] Kryshko, M. (2011). Data-rich DSGE and dynamic factor models.

[57] Kydland, F. E. and E. C. Prescott (1982). Time to build and aggregate fluctuations. *Econometrica: Journal of the Econometric Society*, 1345–1370.

[58] Lewbel, A. (1991). The rank of demand systems: theory and nonparametric estimation. *Econometrica: Journal of the Econometric Society*, 711–730.

[59] Long, J. B. and C. I. Plosser (1987). Sectoral vs. aggregate shocks in the business cycle. *The American Economic Review 77*(2), 333–336.

[60] Long Jr, J. B. and C. I. Plosser (1983). Real business cycles. *The Journal of Political Economy*, 39–69.

[61] Ludvigson, S. and S. Ng (2009). Macro factors in bond risk premia. *Review of Financial Studies 22*(12), 5027–5067.

[62] Ludvigson, S. and S. Ng (2010). *A factor analysis of bond risk premia*, Volume Handbook of Empirical Economics and Finance, pp. 313–372. Chapman and Hall.

[63] Newey, W. and K. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*(3), 703–708.

[64] Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica 77*(5), 1447–1479.

[65] Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics 92*(4), 1004–1016.

[66] Onatski, A. (2011). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *manuscript, University of Cambridge*.

[67] Perron, P. (2006). Dealing with structural breaks. *Palgrave handbook of econometrics 1*, 278–352.

[68] Pesaran, M. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica 74*(4), 967–1012.

[69] Quah, D. and T. J. Sargent (1993). A dynamic index model for large cross sections. In *Business cycles, indicators and forecasting*, pp. 285–310. University of Chicago Press.

[70] Rudebusch, G. and T. Wu (2008). A macro-finance model of the term structure, monetary policy and the economy. *The Economic Journal 118*(530), 906–926.

[71] Sargent, T. (1989). Two models of measurements and the investment accelerator. *The Journal of Political Economy*, 251–287.

[72] Shanken, J. and M. Weinstein (2006). Economic forces and the stock market revisited. *Journal of Empirical Finance 13*(2), 129–144.

[73] Stock, J. and M. Watson (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association 97*(460), 1167–1179.

[74] Stock, J. and M. Watson (2003). Has the business cycle changed and why?

[75] Stock, J. and M. Watson (2009). Forecasting in dynamic factor models subject to structural instability. *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry*, 173–205.

[76] Stock, J. H. and M. W. Watson (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics 20*(2), 147–162.

[77] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

[78] Zhao, P. and B. Yu (2007). On model selection consistency of lasso. *Journal of Machine Learning Research 7*(2), 2541.

[79] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.